Development of New Methods for the (Q)SAR Applicability Domain Assessment: Using Structural Information in a Statistical Study of the Errors in Prediction

Thesis

DEVELOPMENT OF NEW METHODS FOR THE (Q)SAR APPLICABILITY DOMAIN ASSESSMENT USING STRUCTURAL INFORMATION IN A STATISTICAL STUDY OF THE ERRORS IN PREDICTION

Rodolfo GONELLA DIAZA

Thesis submitted for the degree of

Doctor of Philosophy

In

Life and Biomolecular Sciences

at the Open University, UK

IRCCS-Istituto di Ricerche Farmacologiche "Mario Negri"

Milan, Italy

March, 2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Rodolfo Gonella Diaza

Development of new methods for the (Q)SAR applicability domain assessment: using structural information in a statistical study of the errors in predictions

Keywords: QSAR, in silico, computer-based models, applicability domain, predictive toxicology, structural alerts, chemical classes.

Abstract

The main aim of (Q)SAR is to build models to evaluate and predict properties of molecules, such as biological and environmental effects, and physico-chemical properties. These models are built using available experimental data, whose quality and quantity heavily affect their capability of obtaining reliable predictions for new chemicals. A dataset can be viewed as a "sampling" of the whole chemical space, if a sample is too small and / or too homogeneous, the model will inevitably have limitations in the type of chemicals it can predict.

From the point of view of protecting the human health and the environment, it is preferable that a model is able to predict even a small number of chemicals, but with the highest possible reliability. The "coverage" issue can be overcome by integrating results from different models. In this perspective the importance of clearly defining the model's applicability domain is crucial to identify which model is most suitable for each chemical to assess.

The definition of the applicability domain (AD) of (Q)SAR models is still an open research field. Several approaches have been proposed and implemented through years, including the use of structural features such as functional groups and atom-centered fragments. These features have also proven to be useful for an *a priori* definition of AD, making it independent from the specific algorithm chosen to develop the model.

Within this study, the definition of (Q)SAR models' applicability domain has been investigated using structural features of different complexity: thresholds for chemical composition and molecular weight, chemical classes related to commonly well and badly predicted molecules, and statistically-extracted structural fragments to model the error in prediction. In the case studies considered, these approaches improved the AD definition provided by the model developers, supporting their integration within the definition of the models' applicability domain.

Declaration

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Rodolfo Gonella Diaza

Acknowledgements

I would like to thank my parents. You have always respected my decisions and I have always and constantly felt your support, in any aspect of my life. Thank you very much! I love you.

Azzurra, however you have entered in my life only less than two years ago, your strength has fuelled me and your smile and energy have made my days brighter than ever. I love you so much.

Special thanks go to my supervisors: Dr. Emilio Benfenati and Prof. Daniel Neagu for their support, advice and guidance during the entire period of my Ph.D. studies.

Many thanks to Serena Manganelli, Dr. Alessandra Roncagliani, Alberto Manganaro, Giuseppa "Nelly" Raitano and Dr. Nazanin Golbamaki, for your help, collaboration and advices.

I would also like to thank the members of the panel of my final viva, Dr. Judith Madden, Dr. Marco Gobbi and Dr. Valentina Bonetto. I feel that the discussion has been a really constructive moment and it has been a pleasure to meet you.

A big thank goes to some special friends I met during this period at Institute Mario Negri: Michela (ti voglio bene fex), Giacomo (my "brother in arms"), Laura, Marinella, Claudia, Alessia, Max, Anna, Serena, Fabiola, Alice, Giorgio, Nazanin, Azadi, Nelly, and thanks to

all my present and former colleagues and friends for the very special time we spent together.

I would like to thank also some very special friends of mine. Davide, Tazzi, Matt, Dabru and Miguel (playing with you in Odio Su Tela it's been truly amazing and exciting, you all are and will be my brothers forever), Monica, Anna, Francesca, Cristina, Serena, Simona, Cecilia, Grass, Albert, Simone and Mara (I love you all, thanks for everything), Gary, Dero, Sid and Pietro from Node (on the stage we are a true war machine, and I love it!).

A very special and huge thank goes to one of the most important person in my life. We have been friends for 15 years, nearly half of our whole life. What "surprise" me, is that even if we don't meet for months, when we do it is like we have met every day since the day before. You are much more than the sister I never had. Thank you for everything Vale!

Contents

List of Figures

List of Tables

"Don't get set into one form, adapt it and build your own, and let it grow, be like water. Empty your mind, be formless, shapeless — like water. Now you put water in a cup, it becomes the cup; You put water into a bottle it becomes the bottle; You put it in a teapot it becomes the teapot. Now water can flow or it can crash. Be water, my friend."

- Bruce Lee

"...Faith drives me to carry on, and take the road less travelled on..."

- Robb Flynn

Chapter A.
Introduction

1. Background and motivation

Since their introduction in the 19th century, (quantitative) structure-activity relationships ((Q)SAR) models have drawn increasing attention from the scientific community. The underlying idea of this modelling approach is that biological effects and properties of chemicals should be related to their structural properties. Therefore, it should be possible to model this relation using a set of molecules with known structures and experimentally-determined biological effect. Moreover, being based on mathematical and statistical approaches, (Q)SAR models are relatively easy to implement within automated computational tools. This gives the possibility to screen thousands of molecules in a short time and with lower costs, increasing the interest of both industries and regulators, directly or indirectly involved in chemical safety.

As for *in vivo* and *in vitro* models (and every model in general), the results obtained by (Q)SARs are affected by uncertainty. Two main sources of "errors" can be identified: the quality of the data used to build the models, and the use of "wrong model for the wrong molecules". This thesis focuses on the second aspect, which relates to the commonly known difficulty of models to extrapolate reliable results for elements too different from those used to train them. To solve the problem of defining which

molecules can be reliably predicted by a (Q)SAR model is commonly known as the definition of the applicability domain of a model.

In a recent study, the LIFE+ project ANTARES has identified and evaluated 50 models (including commercial and freely-available software) for eight biological properties [56,57,58,59,60]. The method adopted for this study consisted in predicting large dataset of chemicals with known experimental values and evaluate the predictive performance using statistical analysis, such as the "coefficient of determination" (R^2) for models providing continuous numerical results, and the accuracy, sensitivity and specificity for classification models. The most interesting outcomes were obtained while considering the applicability domain information provided with each model, and the training sets used to build these models. (Q)SAR models can generally obtain more reliable results for molecules used to build them compared to "new" chemicals (this was confirmed by ANTARES). The use of applicability domain information proved to give the possibility to improve the identification of chemicals predicted with higher reliability also for the new chemicals.

2. Problem description

As introduced above, (Q)SAR models are built using mathematical and statistical approaches. Initially, linear modelling approaches were used, making it relatively easy to define their applicability domain. For example, range-based methods describe the applicability domain using the ranges of the variables used by the model, and that of the experimental responses for the molecules of the training set. If the variables calculated

for the a new molecule or the prediction obtained are out of the ranges defined by the training set, the molecule can be excluded by the applicability domain.

More complex modelling approaches are used nowadays to build (Q)SAR models (e.g. Multiple Linear Regression, Partial Least Squares, Neural Networks, etc.). Range-based approaches can still be applied, however they have been joined by other methods (e.g. geometric, distance-based, probabilistic, etc.). The ANTARES evaluation has demonstrated their usefulness, showing however that we are still far from the perfect discrimination between reliable and unreliable predictions.

As described in the next chapters, most of the available methods are based on the evaluation of the molecular descriptors used to build the (Q)SAR model, or evaluate the structural similarity between the target chemical and the model's training set. Some successful attempts to "join" these two aspects have been described, by defining the applicability domain using atom-centered fragments or considering chemical classes (defined by the presence of functional groups).

3. Thesis aims

The main aim of the research activities reported in this thesis was to study the use of the above mentioned structural features, for the definition of (Q)SAR models' applicability domain. In particular, three research directions were investigated:

- The influence of simple properties distribution on the reliability of model's predictions,

- The capability of models to predict different chemical classes with different accuracies, and
- The possibility to model the error in prediction using the same techniques used to build (Q)SAR model, considering it as an endpoint.

Moreover, an ambitious goal was also set: using these approaches to study the applicability domain *a priori*, making its definition endpoint-dependent rather than model-dependent.

4. Methodology

To investigate the possibilities described above, considering that several types of (Q)SAR models can be built, depending on the type of endpoint and approach used, three endpoints were considered as case studies: bioconcentration factor (BCF), mutagenicity, and oral rat acute toxicity. For each endpoint a dataset of molecules provided with experimental values was available thanks to the ANTARES project: 860 molecules were available for BCF, 7420 for acute toxicity, and 6065 for mutagenicity. Nine (Q)SAR models, developed using different types of data and approaches, were also selected to assess how the proposed solutions could affect different type of models.

Several statistical methodologies were adopted to achieve the research aims, including also the visualisation of graphical plots of the data and results:

- Histograms were applied for the visualisation of the distribution of the predictive capabilities among classes of molecules, defined using thresholds on simple

properties (e.g. the molecular weight). These histograms were then used to set thresholds for the definition of the applicability domain;

- The occurrences of several predefined functional groups were analysed using a freely-available tool (istChemFeat). This software also calculates the distribution of a target property among the identified chemical classes (e.g. the percentage of mutagenic chemicals within a chemical class);
- The possibility to model the error in prediction was investigated using another freely-available software (SARpy) which builds a library of molecular fragments starting from a training set, and extracts those more relevant for the target endpoint (using the likelihood calculation).

The applicability domain definitions were used with predictions from different (Q)SAR models, to assess their ability to discriminate between reliable and unreliable ones. The datasets used for the study were split to discriminate between the model performance for molecules belonging to its training set, and those (more interesting) for "new" chemicals.

5. Thesis structure

The thesis is arranged in eight chapters (grouped in three main parts), and two annexes for the supplementary data:

Part I: is the introductory part, containing the theoretical introductions and the literature review. It is organized in three chapters:

- Chapter B – contains the introduction to (Q)SAR models, including the theoretical basis, the available modelling approaches and an introduction to their evaluation;
- Chapter C – introduces the concept of applicability domain of (Q)SAR models and gives an overview of the approaches currently used for its definition;
- Chapter D – contains examples of successful use of structural features for the definition of the applicability domain.

Part II: is the complete explanation of the work done and the report of the results obtained. It is organized in two chapters:

- Chapter E – is the explanation of the methods used and the framework of the research activities. This chapter contains the explanation of all the data, endpoint, models and software used as case studies;
- Chapter F – presents all the results obtained and is organized by the research directions investigated.

Part of the results and methods presented in these chapters have been published in: Gonella Diaza R, Manganelli S, Esposito A, Roncaglioni A, Manganaro A, Benfenati E. Comparison of in silico tools for evaluating rat oral acute toxicity. *SAR QSAR Environ Res.* 2015 Jan;26(1):1-27.

Moreover, an oral presentation entitled "Applicability domain for mutagenicity models: an *a priori* approach, based on chemical classes" was given at the 16th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences (QSAR2014), June 17th 2014, Milan, Italy

Part III: is the conclusive part of the thesis, organized in two chapters:

- Chapter G – is the complete discussion and comparison of the results obtained from the three main research activities;
- Chapter H – contains the conclusive notes and future perspective.

Part I – Background: Applicability domain of (Q)SAR models

Chapter B.

(Quantitative) Structure-Activity Relationship models

6. Determination of chemical properties: methods for the assessment of chemical safety

In the contemporary society, a huge number of chemicals are widely used in a variety of human activities, including food colouring and preservatives, drugs, pesticides and many others. According to what was reported by the American Chemical Society's Chemical Abstract Service (CAS), more than ninety million organic chemicals have been registered so far [1]. Chemicals present in the environment and the food may interact with biological systems, posing a high risk to the environment itself and to humans. For this reason, the determination of the toxic activity of chemicals has become more and more a primary objective in our society.

The interaction of a chemical with a biological organism may lead to a visible effect (e.g. a specific disease) but the biological mechanisms underlying this effect are often unknown or only partially described. This poor knowledge makes it nearly impossible to have a clear idea of the possible effects of a chemical; thus toxicity must be studied experimentally.

It is obviously not possible to obtain all the experimental information from tests on humans. In drug discovery, for example, human testing (called clinical trials) is the very

last step of the development. Three types of approaches can be used to study the biological activity (including toxicity) of chemicals: *in vivo* (involving animal experiments), *in vitro* (e.g. using tissue culture cells), and *in silico* (also referred as non-testing methods, involving computer-based screening).

In silico approaches have drawn more and more attention of chemical and pharmaceutical industries, academia and regulatory bodies, due to their lower costs both in terms of time and money. Moreover, animal testing is becoming unacceptable among a growing number of people. For these reasons, the scientific community and the industrial world have started to use and develop *in silico* models. Virtual screening, for example, is currently widely used in the first steps of drug design. In the last few years, computer-based models have become acceptable also from the regulatory point of view. For example, the Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) clearly states that results obtained from *in silico* models can be used for the registration of chemicals and provides general guidelines for the acceptability of values obtained with these methods [2]. The requirements established by REACH (and other regulations) will be described in Chapter C.

7. Computer-based models to predict biological activity

7.1 Biological effects depend on the chemical structure

The biological activity of molecules depends on the possible physico-chemical interactions they can establish with a large number of enzymes, receptors and other macro molecules at the cellular level. These interactions trigger specific mechanisms, which finally lead to a biological response.

Three-dimensional conformation and size play a key role in the possibility for a molecule to interact with the correct target. From this perspective, the molecular structure seems to be strictly related to the biological effect of chemicals. Moreover, structurally similar molecules should activate the same biological pathways, leading to the same effects.

7.2 Similar molecules share similar effects

Johnson and Maggiora in the early nineties introduced the similarity property principle (SPP) [3], which stated that similar compounds should have similar properties, clearly referring to biological activities. Higher similarity between molecules seems also to lead to an increased similarity of their effect [4]. From another point of view, a strong relationship between molecular similarity and biological effects has been demonstrated using common substructures to group similar molecules [5].

The introduction of the SPP and the subsequent confirmation of the relation between structure and activity [3,4,5], have led to the development of a number of

similarity-based methods to determine biological properties. These approaches have become popular for example in pharmaceutical industry and medicinal chemistry [5,6,7,8]. The drug discovery process implies the design and testing of large libraries of molecules and can take advantage of a fast method to obtain *a priori* information on the activity and toxicity of drug candidates. If the similarity between molecules can be parametrized and modelled using mathematical relationships, such models can then be implemented as computer programs, considerably speeding the screening of large libraries of chemicals.

7.3 Using computational tools to relate structure similarity and effects

To build mathematical relationships between chemical structure and a particular biological effect, it is necessary to represent the structure using numerical variables (either discrete or continuous). The underlying idea is that, if it is possible to find a relation between structure similarity and a particular biological effect, this similarity will be also codified in the numerical variables, which than could be used to obtain predictive equations. These variables are generally called "molecular descriptors". A number of methods have been developed in the last decades to calculate molecular descriptors, ranging from simple calculation (such as the molecular weight) to complex molecular fingerprints [19,20,21]. Using these descriptors can help in transforming the study of the biological effect of a chemical in a data mining problem, which can be automated and implemented in computer-based programs.

The data mining algorithms search for a mathematical relationship between a set of relevant molecular descriptors and the so called "endpoint of interest", which is the biological effect to model. This means that, to develop mathematical predictive models for a certain endpoint, it is necessary to have a set of molecules with known structures and experimentally determined values of the property to model. Moreover, the reliability of the model obtained will depend on the number of molecules available and the quality of the experimental data. The mathematical relationship obtained using this approach can be usually implemented in software, which can be used to evaluate the endpoint for new molecules, for which experimental values are not available.

8. The Structure-Activity Relationship approach to build computer-based models

Structure-Activity Relationship (SAR) and Quantitative Structure-Activity Relationship (QSAR) approaches represent families of mathematical and statistical methods to computationally find the relation between the structure of similar molecules (represented using the molecular descriptors) and the endpoint of interest. As the name suggests, the main difference between SAR and QSAR is related to the modelled property: SAR models are developed for biological activities usually represented by categories (e.g. toxic or non-toxic); on the other hand QSAR deal with properties represented by continuous values (e.g. the bioconcentration factor). The term (Q)SAR is commonly used to generally refer to both families.

Historically, the first application of a (Q)SAR model dates back in the 19th century. Early studies highlighted correlation between toxicity of organic chemicals and their

water solubility [9] and lipophilicity [9,10]. After these initial results, the (Q)SAR approach drew some attention across the scientific community, however these methods started to be accepted and used mainly thanks to the pioneer work of Corwin Hansch [11,12], considered nowadays the founder of modern (Q)SAR modelling.

(Q)SAR models find applications among scientific communities involved in different matters (e.g. toxicology, environmental effects and pharmaceutical research). Depending on their applications, (Q)SAR models are built based on different approaches. Models built for environmental related endpoints (e.g. ecotoxicity) are commonly based on the partition coefficient between octanol and water (LogP), as well as constitutional descriptors (e.g. the molecular weight) and electronic features (e.g. "eHOMO, the Highest Occupied Molecular Orbital") [13].

Another approach for building predictive models consists in the implementation of rules, based on the presence of structural fragments (usually called structural alerts) correlated to the effect to model. In 2008, for example, the European Joint Research Centre (JRC) published a Scientific and Technical report, presenting a predictive software (ToxTree) which included models for predicting mutagenicity and carcinogenicity, based on the structural alerts included in the Benigni-Bossa ruleset for mutagenicity and carcinogenicity [14]. The main differences between the two approaches is that in the first case, the descriptors are chosen using statistical methods, whereas the Benigni-Bossa rules derived from experimental evidence. In this case predictive models are called "knowledge based". Another important difference between these two examples, is the type of endpoint modelled. Environmental toxicity is usually measured with

continuous values, such as the Median Lethal Concentration (LC50), which is the chemical concentration that is expected to kill 50% of a group of organisms. This concentration can be correlated with molecular descriptors (such as the LogP) using mathematical equations. On the other hand, mutagenicity and carcinogenicity are usually expressed as a binary concept: toxic or not. In this case, the hypothesis is that the presence of a single toxic-related molecular fragments, could be enough to make the chemical toxic. In the first case the endpoint is quantitative, we can therefore speak of QSAR models. On the other hand, mutagenicity and carcinogenicity models based on structural alerts are "qualitative" and we can speak of SAR models.

8.1 The importance of experimental data

To build a (Q)SAR model for predicting the biological effects of chemicals, it is necessary to have experimental data of the effect produced by a set of molecules. Animal and *in vitro* tests are the main sources of experimental values. It is however fundamental to understand that both animals and micro-organisms used to test the effects of chemicals are models. This means that all the experimental values are associated with an uncertain value. For example, in the case of environmental-related effects and toxicity, the reported experimental uncertain for the bioconcentration factor can be up to 0.75 in Log unit [15]. Also in the case of the evaluation of human toxicity, a certain degree of uncertainty is accepted. The reproducibility of the Ames mutagenicity test, which is a quite simple model using bacteria, is about 85% [16].

The importance of uncertainty information related to the experimental data to use is crucial. Let's consider for instance the LC50 parameter used in ecotoxicology: this parameter indicates the dose able to kill half of the animals of the test population. Why this happens and the underlying mechanism(s) are often largely unknown. This kind of probabilistic information is perfectly useful within, for example, the risk assessment framework. When experimental data are used to build (Q)SAR models, the uncertainty should be clearly defined, as it will also affect the uncertainty of the model. The uncertainty of the final model cannot be inferior to that of the input data, and it is suspicious to obtain values predicted with a precision superior to that of the experimental laboratory model.

While building (Q)SAR models, the most important sources of experimental data are on line databases of chemicals, and peer-reviewed publications. It has to be considered that errors and values with high uncertainty can be found both in databases and the literature [17,18]. For this reason it is important, when possible, to use multiple sources of data, in order to compare and integrate experimental values obtained by different research groups and laboratories.

8.2 Structural information for models: descriptors and fragments

Many algorithms have been used and implemented to calculate molecular descriptors from the structural information of molecules [19,20,21]. Some of them also include libraries of pre-codified structural fragments, like functional groups and atom-centered fragments. Besides these pre-codified fragments, other approaches have been

developed to analyse a dataset of molecules and build new fragments which correlate with the endpoint to model [22].

8.2.1 Different descriptors to codify for different structural aspects

A large number of molecular descriptors have been developed in the last decades, describing different aspects of the structural information of molecules [23]. Constitutional descriptors include simple properties like molecular weight, number of atoms present in a molecule (for instance number of chlorine atoms), number of double bonds, etc. Topological descriptors contain information about the number and type of bonds between atoms, and can be used to represent the ramification of the molecules. Certain descriptors consider the electronic charge and polarity of the atoms in the molecules. Even more complex descriptors are able to represent the molecular orbitals. For example eHOMO and eLUMO refers to the energy of the highest occupied and lowest unoccupied molecular orbitals. Physico-chemical descriptors include parameters such as the partition coefficient between octanol and water (LogP) and lipophilicity.

8.2.2 Chemical structure codification to calculate descriptors

Molecular descriptors are usually calculated using specific software. This means that the chemical structure has to be represented in a suitable way. Currently the most common formats are InChI [24], SMILES [25] and MDL molformat [26]. Commonly, the last two formats are used by software for either descriptor calculation and fragments comparison or extraction. SMILES is probably the easiest and most simplified way to

represent molecules. Each molecule, in fact, can be codified in a single string of characters. The main problem with this formalism is that different algorithms can generate different SMILES for the same molecule. This must be taken into account while creating and using models, making sure to use always the same formalism for all the molecules.

8.3 The modelling algorithms: classifiers and regression models

Beside the development of new molecular descriptors, also more and more advanced and powerful modelling algorithms have been introduced. While the first (Q)SAR models were developed as simple linear combination of molecular descriptors, the last decades have seen the development of algorithms such as neural network, fuzzy logic and data mining. These methods are also referred as "pattern recognition methods" because their aim is to devise algorithms that could learn to distinguish patterns in a data set. Using these advanced mathematical and statistical approaches, predictive models based on non-linear correlation have been generated [27,28].

The algorithms used to develop (Q)SAR models can be classified on the basis of the type of endpoint modelled. Regression methods are used to develop QSAR models, which provide a quantitative evaluation of the biological effect. On the other hand, classification approaches are used to develop SAR models which are useful to categorize chemicals. Regression approaches can also be used indirectly to classify molecules. The acute toxicity effect, for example, is usually represented using the median lethal dose (LD50), which is a continuous value. However, the predicted LD50 values can be used to

categorize the chemicals, on the basis of the Acute Toxicity Estimate (ATE) thresholds, defined within the European CLP regulation.

Another important distinction among the large variety of mathematical and statistical algorithms, used in the (Q)SAR development, consists in how the experimental values of the endpoint are used. The so-called "supervised" methods (e.g. Multiple Linear Regression, Discriminant Analysis, Partial Least Squares, Classification and Regression Trees, Neural Networks, etc.) utilize the biological effect information to select the molecular descriptors and build the relationships between them and the effect. On the other hand, the "unsupervised" methods (e.g. Principal Component Analysis, Cluster Analysis, k-Nearest Neighbours, Nonlinear Mapping, etc.) do not use the experimental data, and only search for patterns in the descriptor data. The advantage of unsupervised learning is the lower likelihood of chance effects, due to the fact that the algorithm does not try to fit a model.

8.4 How to build a (Q)SAR model

As explained above, the entire development of a reliable (Q)SAR model depends on the quality and adequacy of the available experimental data. It is therefore very important to check the chemicals of the dataset to use [29]. This initial step is called "data curation" and includes several steps, such as:

- deleting inorganic and organometallic compounds, counterions, salts and mixtures;

- deleting duplicate chemicals, paying attention to the experimental data associated to them;
- checking the validity of the structures and the correct ring aromatization;
- normalizing specific chemotypes;
- checking the consistency of tautomeric forms.

If the structural and experimental data have been taken from multiple sources, it is also crucial to compare the different datasets in order to find and resolve possible overlaps. As already mentioned, it is common that the same chemical will be associated with different experimental values in different datasets. The decision on how to integrate different data for the same chemical should not be underestimated and the simple use of statistical methods, such as the mean or median calculation, may not be appropriate or sufficient.

Once the dataset has been selected and checked, the next step is translating the structural information into algorithm-readable molecular descriptors, using one or more software available. Conceptually, a table of all the molecules is built up. Each chemical is associated with the molecular descriptors and the experimental value for the property to model. From the mathematical point of view, the molecular descriptors represent the independent variables x (the input) of the model, and the experimental values are the dependent variable y (the output).

Since thousands of molecular descriptors can be generated, it is necessary to select only those necessary to best explain the structure-activity correlation. Models with an unnecessary high degree of complexity may be affected by the so-called "over fitting"

problem [30,31]. An uncritical use of the powerful mathematical modelling tools may in fact lead to a model that is "too trained" in evaluating the molecules of its training set and it's not able to reliably predict new chemicals.

The Principal Component Analysis (PCA) is a method usually applied to reduce the number of variables to consider. This approach performs a linear combination of the original variables to create a smaller set of new ones, which explain most of the variability of the dataset. Variable elimination is another family of techniques by which unhelpful or unnecessary variables are removed from a data set [32,33]. Both PCA and variable elimination are usually applied to the whole set of molecular descriptors calculated, before the very model building step. Even after eliminating unnecessary variables, there may still be many variables to choose from. In this case variable selection is used, which can choose descriptors that will be useful for mathematical modelling and will lead to a model able to predict new compounds. There are many diverse procedures for variable selection and some are built in to the process of model building, such as forward stepping multiple regression [34].

The selection of the correct molecular descriptors to be used in a predictive model can be also made on the basis of expert knowledge. This is the case of SAR models, based on structural alerts. For example, the genotoxicity effects of nitroaromatic compounds and aromatic amines are well known [35,36]; thus, the presence of these structural groups can be used to classify a molecule as genotoxic [14].

8.5 The validation of a (Q)SAR model

The major problem with (Q)SAR models, and predictive models in general, is the reliability of predictions obtained on "new chemicals", which were not included in the training set. Therefore, models have to be validated. Although it may seems obvious, this concept is clearly stated within regulations. Indeed, as will be described later, the statistical validation of (Q)SAR models is one of the five requirements described within the OECD principles for (Q)SAR [37].

While goodness-of-fit and robustness refer to the performance on the molecules of the training set (internal performance), the predictivity has been introduced to precisely assure that the model can obtain reliable predictions on new chemicals (external validation).

Many statistical tools have been introduced for internal validation [38], and they can be classified in three main groups:

- Cross validation approaches leave one or more molecules out of the training set, build the model, and evaluate its performance on the left out molecules. In the first case we speak of Leave One Out Cross Validation (LOO CV), whereas in case of many molecules left out (usually the 10 or 20% of the dataset) we speak of Leave More Out Cross Validation (LMO CV).
- Bootstrapping simulates what happens by randomly re-sampling the data set with n objects. Multiple n-dimensional groups are generated by eliminating some of the compounds. Each group is then used to build the model, which will be then

evaluated on the excluded chemicals. The estimation of the model predictivity is then obtained by the average of all the evaluations.

- Y-scrambling is used to evaluate the probability that the model predictions have been obtained only by chance. While keeping the molecular descriptors in the correct order, the experimental data are randomly reassigned and the model performance are evaluated at each iteration.

The evaluation of the internal performance is important and useful while building the model, but it has been clearly stressed that it is not enough to consider a (Q)SAR model as reliable [39]. The main problem is that even the left out molecules (in cross validation and bootstrapping) cannot be considered external, since they are used to build the best correlation. An external dataset entirely composed of new molecules is therefore necessary to evaluate the predictivity of a model.

8.5.1 Evaluation of a classifier

As already described, SAR models evaluate biological properties qualitatively. The molecules are simply classified by their probability of showing or not a certain biological effect. The typical approach to evaluate these models is using the Cooper statistic. In the case of binary classification of toxic (positive) and non-toxic (negative) molecules, the predictions obtained can be grouped in four classes:

- True positive (TP): toxic molecules predicted as toxic;
- True negative (TN): non-toxic molecules predicted as non-toxic;
- False positive (FP): non-toxic molecules predicted as toxic;

- False negative (FN): toxic molecules predicted as non-toxic.

These classes are used to calculate three main statistical parameters for model evaluation:

- Accuracy (A) is the measure of the correctness of prediction. This parameter gives a general evaluation of the errors made and is defined as the ratio between the compounds correctly predicted and the total number of compounds.

$$A = \frac{(TP + TN)}{Total} \quad [1]$$

- Sensitivity (S) is the measure of the positive compounds correctly predicted and is defined as the ratio between TP and the total number of positive (TP + FN).

$$S = \frac{TP}{(TP + FN)} \quad [2]$$

- Specificity (SP) is the measure of the negative compounds correctly predicted and is calculated as the ratio between TN and the total number of negative (TN + FP).

$$SP = \frac{TN}{(TN + FP)} \quad [3]$$

In the ideal case, a model should have high values for all these three parameters. However, while performing the predictivity evaluation, FP and FN could have a different importance depending on the scope of the prediction. For example, the regulatory point of view is usually "conservative", a model affected by a high ratio of FN (low sensitivity) is not well accepted since it would classify hazardous chemicals as safe.

8.5.2 Evaluation of a regression model

Regression approaches are used to develop QSAR models for the evaluation of biological effects associated with a numerical values (e.g. LogP, bioconcentration factor, etc.). The most used parameter for the evaluation of a regression model is the coefficient of determination (R^2), which measures how well data fit a statistical model.

$$R^2 = 1 - \frac{\sum_i (y_, - \bar{y})^2}{\sum_i (y_, - f_i)^2}$$

Where $y_,$ is the experimental value of the molecule i, $\bar{y}$ is the mean value of the experimental values of all the molecules considered, and f_i is the predicted value of the molecule I. R^2 values ranges from 0 (bad correlation) to 1 (perfect correlation).

R^2 is usually calculated for both internal validation and external validation. In the first case the correlation is calculated between experimental and predicted values of the training set molecules, and gives important information about the goodness-of-fit and the robustness of the model. R^2 is also used within LOO, LMO, bootstrapping and y-scrambling methods, both to validate and to select the best model, during the development steps. Finally, R^2 is used for the external validation of the QSAR model, to evaluate the predictivity. In this case, R^2 is calculated on the predictions obtained for new molecules, comprising the validation (or test) set.

Chapter C.

Applicability domain of (Q)SAR models

9. When the similarity principle fails

As described in the previous chapter, (Q)SAR modelling is historically based on the similarity property principle (SPP), which states that similar chemicals should share similar effects. The core problem of this assumption is a clear and consistent definition of similarity between molecules [40]. Small modifications in the chemical structures may lead to completely different effects, creating "discontinuity" in (Q)SAR models. These so-called "activity cliffs" effect represents a major limitation of the SPP assumption [41,42,43]. The definition of similarity, especially if based on individual perspective, can therefore highly influence the predictivity of (Q)SAR models.

Several aspects and definitions of similarity between molecules further complicate the basis of (Q)SAR modelling. Chemical and molecular similarity are often used with the same meaning, however they are based on different criteria. Chemical similarity is primarily a physico-chemical comparison, based on parameters such as molecular weight, solubility, LogP, electron densities, etc. On the other hand, molecular similarity is based on structural features (functional groups, ring systems, substructures, etc.). Another important difference resides in the molecular representation. Properties and effects of chemicals result from the interaction they can make with others, which

depend on their three-dimensional conformation. However, given the uncertainties associated with identifying the "correct" 3D conformation that molecules assume during the chemical and biological interactions, the application of bi-dimensional parameters is often preferred while calculating similarity. However some important bioactive-related information are almost inevitably lost with the 3D to 2D approximation, bi-dimensional approaches have proven to be more robust in (Q)SAR modelling [44,45].

Starting from an opposite point of view, the biological similarity can be introduced for the comparison of chemicals. The usual molecular descriptors are replaced by the activities of chemicals against several biological targets (usually proteins), and the chemicals are compared using specific pairwise similarity approaches, which do not account for structural features [46,47].

The concepts of similarity described so far are all based on the comparison of the chemicals as a whole (global similarity); the structure, property or biological effect refer to the entire structure of the molecule considered. In contrast to these definitions, the local similarity between two or more compounds can be evaluated on the basis of a small subset of atoms. This approach is often used for example in drug design, to search for a pharmacophore [48]. In this case the similarity algorithm only focuses on the subset of atoms of interest (the pharmacophore), whereas the rest of the molecule is not considered. The base assumption of this approach is that if two molecules share similar pharmacophore elements, they will very probably share also the same activity [49].

The computational pharmacophore model built using local similarity approach resemble the structural fragments-based SAR models, in terms of how they determine

the activity of a molecule. Both models base their assessment on the presence of a particular structural moiety in the molecule. However, as aforementioned, chemicals may show similar biological effect even if they do not "seem" similar. This means that the absence of target substructure does not mean that the chemical does not share the same biological effect. Thus, if a chemical does not have the pharmacophore (or the structural alert), this does not mean that it will not share the same biological effect, it only means that the model is not able to predict its activity.

10. What is the applicability domain?

10.1 Domain of a predictive model

As stated by Brooks at al. in 1988 [50]:

> *"A primary application of regression analysis is prediction. We call the region in which prediction is valid the domain of the model. The definition of the domain is important because predictions made outside the domain may be unacceptably different from the true responses".*

When predicting a new point, which was not used during the regression calculation, it is important to understand if the prediction results from an interpolation or is an extrapolation. In this second case the point is outside the domain and may be misleading. Considering the very simple case of a regression model depending on only one predictor variable, the domain can be defined as the range of this variable for the set of experimental data used.

10.2 Applicability domain of (Q)SAR models

The concept of applicability domain (AD) of a model applies also to both quantitative and qualitative (Q)SAR models. QSAR models are usually composed by many molecular descriptors, making the determination of the AD more difficult than those mentioned in the previous paragraph [50]. The case of SAR models is different; since they mostly base their prediction on the presence of particular molecular fragments, the AD is theoretically intrinsic in the model itself. Molecules which do not have any of the fragments included in the model cannot be predicted. Thus, they should be outside the model's applicability domain.

A good definition of the applicability domain of (Q)SAR models has been given in 2005 by Netzeva et al. in the report of the 52nd ECVAM Workshop [51]:

"The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability"

The chemical structure could be represented by physico-chemical and/or fragmental information whereas the response could be any physico-chemical, biological or environmental effect predicted by the (Q)SAR model.

10.3 The importance of a defined Applicability Domain

The importance of a clear definition of the domain of application of predictive models is clearly stated within the definitions given both by Brooks et al. and Netzeva et al. [50,51]. When a cause-effect relationship is modelled, the values predicted are not useful and enough *per se*. The reliability of these values must be evaluated in order to provide the

end-users with the means to decide if they can trust the results, and to what extent. Therefore, for (Q)SAR models to be used as predictive tools, their applicability domain must be defined [52].

Considering the current applications of (Q)SAR models as instruments for evaluating biological and environmental effects, such as the toxicity, the clear definition of their AD has become even more important and has been also stressed within regulatory frameworks. On February 2003, with the so-called seventh-amendment to the Council Directive 76/768/EEC (which was aimed at regulating cosmetics products in Europe), the European Community (EC) has imposed an animal testing ban for cosmetic products in favour of alternative methods, including (Q)SAR models. Within the 2006's REACH regulation (Registration, Evaluation, Authorisation and restriction of Chemical substances), EC has introduced the duty of compiling dossiers with complete physico-chemical and (eco)toxicological information for chemical substances circulating in Europe, depending on their tonnage. In order to avoid a high usage of animal testing, REACH also states that these tests have to be replaced by alternative tests, if available. Foreseeing the usage of (Q)SAR methods for the evaluation of possible threats to human health and environment, the REACH regulation has been provided with a series of requirements that must be met for the acceptance of results from (Q)SAR models. As outlined in the annex XI of the legislation:

> *Results obtained from valid qualitative or quantitative structure-activity relationship models ((Q)SARs) may indicate the presence or absence of a certain dangerous property. Results of (Q)SARs may be used instead of testing when the following conditions are met:*

1. *results are derived from a (Q)SAR model whose scientific validity has been established,*
2. *the substance falls within the applicability domain of the (Q)SAR model,*
3. *results are adequate for the purpose of classification and labelling and/or risk assessment, and*
4. *adequate and reliable documentation of the applied method is provided.*

As already reported in the previous chapter, also the Organization for Economic Co-operation and Development (OECD) has developed the "OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship model". With these principles, OECD states that:

To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1. *a defined endpoint,*
2. *an unambiguous algorithm,*
3. *a defined domain of applicability,*
4. *appropriate measures of goodness-of-fit, robustness and predictivity,*
5. *a mechanistic interpretation, if possible.*

The definition of a model's AD is also important in the earlier steps of model building and validation. Since the interpolated estimates, in case of continuous properties, are considered statistically more reliable than extrapolated ones, it results that the applicability domain is strictly related to the information present in the dataset used during the model building step [53]. This concept is also clearly stated within the report

of the Setubal Workshop organized by the European Chemical Industry Council (CEFIC) in 2002 [54]:

> *"The AD of a (Q)SAR is the physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The AD of a (Q)SAR should be described in terms of the most relevant parameters i.e. usually those that are descriptors of the model. Ideally, the (Q)SAR should only be used to make predictions within that domain by interpolation not extrapolation."*

The importance and usefulness of a clearly defined AD has been also highlighted within recent evaluations of (Q)SAR models available for REACH-related relevant endpoints, performed within the LIFE+ project ANTARES [55,56,57,58,59,60]. Within these studies, the performance of (Q)SAR models have been evaluated using available datasets. The predictivity of each model has been tested in different conditions: considering the whole dataset, comparing the performance for the molecule present or absent in the dataset used to build the model, and considering the AD available information. The built-in applicability domain tools have proven to be useful to increase the models' performance.

10.4 How to define the Applicability Domain of a model

The definition of the applicability domain of a model could be viewed as an answer to "given a training set, what are the other molecules for which the trained models can be used to obtain a reliable prediction of their properties of interest?"

Several methods have been proposed for the (Q)SARs Applicability Domain assessment over the years. Each method is based on different hypothesis and has its own limitations. Moreover, the AD of a model depends on the model itself and on the algorithm on which is based.

For (Q)SAR models developed applying mathematical and statistical approaches to a training set of chemicals with known experimental information, the applicability domain assessment can be mainly based on the comparison between the values of molecular descriptors of chemicals used as test (or validation) set and the range of the same descriptors calculated for the compounds used during the model building phase. This comparison could be performed using different methods, based on descriptor ranges, new chemical / training set chemicals distance, etc. [51]. Most of these methods, for example, have been implemented within the AMBIT software for chemoinformatic data management [61,62].

Another class of proposed methods for the assessment of AD for statistically developed (Q)SARs is generally related to the structure similarities. The structure of chemicals used to build the model could be split in small *atom-centered fragments* in order to compile a list of all the fragments present on the training set. In order to verify if a new molecule is part of the AD of the model developed, it will also be split in atom-centered fragments, and the resulting list will be compared to that compiled for the training set [63].

Knowledge-based systems represent a completely different approach in modelling biological effects, which are based on the knowledge of expert toxicologists. As

described in the previous chapter, these approaches are mainly based on the identification of structural features associated with the biological effect. These sets of structural alerts can then be implemented into computer software, which are called expert systems [14,64]. The AD of these SAR models, as already said, can be easily based on the same fragments that define the model itself. However, this could not be enough, for example because of other structural properties (e.g. other substructures) which could act as modulators of the structural alerts. For example, in their rulebase for mutagenicity and carcinogenicity, Benigni et al. not only listed structural alerts related to toxicity, in some cases exceptions have been introduced (e.g. the presence of a sulfonic acid group on the same ring of the nitro group seems to decrease the well-known nitroaromatic-related toxicity) [14].

10.4.1 Range-based approaches

Descriptor ranges

The easiest way to describe the applicability domain of a model is to consider the n-dimensional hyper-rectangle defined by the n descriptors composing the model. In this case, a new chemical is considered as being within the AD if all its n descriptor values are comprised in their correspondent descriptor ranges. A major weakness in this approach is that it does not consider areas of the hyper-rectangle, which were poorly described by the model training set.

Principal components ranges

This method is quite similar to the previous one; the main difference is that in this case the ranges considered are not that of the descriptors. The first step is the Principal

Component Analysis (PCA) of the descriptors, which consists of centering the data about the standard mean and then extracting the so-called eigenvalues and eigenvectors of the covariance matrix of the transformed data. The new vectors are characterized by being aligned with the directions of greatest variations in the data set. Principal components ranges utilize these vectors instead of the original descriptor to define the hyper-rectangle identifying the applicability domain; empty spaces are also present but the volume enclosed will be less empty than the original descriptor range.

TOPKAT Optimal Prediction Space

The Optimal Prediction Space (OPS) included in the TOPKAT software use a variation of the standard PCA analysis [65]. The data is centred on the average of each parameter range. The new OPS coordinates are then obtained in same way as eigenvectors and eigenvalues. The boundaries of the hyper-rectangle are here defined by the minimum and maximum values of the OPS vectors. In order to try to face the problem of poorly described areas in the hyper-rectangle, the Property Sensitive object Similarity (PSS) between the test molecule and the training set can be calculated and used to assess the confidence of the prediction.

10.4.2 Geometric methods

The coverage of n-dimensional set can be empirically determined by calculating the convex hull, which is the smallest convex area that contains the original dataset. As for range-based methods, the convex hull is also characterized by regions with a high

density of data points and region where the data are sparse [51]. More sophisticated methods have been developed to face this problem.

Calculation of the Convex Hull is a geometry problem [66]. Efficient algorithms for this calculation can be found for two and three dimensions. The complexity of the problem increases both increasing the number of dimensions and data points (descriptors and chemical substances). This approach, unlike those belonging to the ranged-based one, does not consider the distribution of the data but only analyses the boundary of the data set.

10.4.3 Distance based methods

Distance-based approaches are generally based on the calculation of the distance between the test molecule and the training set of the model. This distance can be calculated between the query molecule and the training set mean, as the average or maximum distance between the query and all the molecules in TS, etc. The threshold to utilize for deciding whether a chemical is out or in the AD has to be chosen by the user.

Several algorithms for distance calculation exist, but three have proven to be more efficient in (Q)SAR: Euclidean, Mahalanobis and city-block distance methods [67]. Each of these three approaches are based on the concept of Distance Matrix (D), which is a square NxN matrix, where N is the number of data points. The general d_{ij} element of the matrix is the distance between the points i and j; this distance can be calculated using one of the three above mentioned methods. Giving a dataset composed by a number p

of molecules each represented by m molecular descriptors, the distance between two molecules (d_{ij}) can be calculated, using the Euclidean approach, in the following way:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

Where x_{ip} and x_{jp} are the values of the p^{th} molecular descriptor calculated for molecules i and j respectively.

The Mahalanobis distance is a weighted version of Euclidean distance [68]; this means that each difference $x_{ip} - x_{jp}$ is multiplied by a factor which considers the importance of the descriptor in the model.

$$d(i,j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \cdots + w_p(x_{ip} - x_{jp})^2}$$

Where w_p is the weight assigned based on the importance of the p^{th} descriptor. This weight-based approach seems to provide a better definition of AD [68,69,70].

The city-block distance is calculated as the sum of absolute differences between the corresponding descriptors of different molecules:

$$d(i,j) = \sum_{m=1}^{p} |x_{im} - x_{jm}|$$

10.4.4 Hotelling T^2 and leverage methods

Hotelling T^2 test and leverage [71,72] are particular cases of distance-based methods, which assume a normal distribution of the data [73]. Hotelling T^2 is a multivariate Student's *t*-test method, whereas leverage is based on the hat matrix $H = (X(X'X)^{-1}X')$, whose diagonal element (h[ii]) represents the distance between the X value

for the i^{th} observation and the means of all X values, indicating if X values may be outliers [74,75]. Both methods use a covariance matrix to correct for colinear descriptors.

Hotelling T^2 and leverage measure the distance between each observed value and the centre of all the observations. The decision of considering an observation as outlier is then taken using cut-off thresholds. In general, higher values mean that the observations are distant from the centre and have to be considered outliers. However, if a high leverage point fits the model, it is called "good high leverage points" or good influence point and acts to stabilise the model and makes it more precise.

10.4.5 Probability based methods

In order to estimate the probability density of a data set, two types of methods can be utilised: parametric and non-parametric approaches. The difference between these two types of methods is that those belonging to the first group assume that the probability function has a standard shape (e.g., Gaussian) whereas the others do not make any initial assumption and estimate the function from the dataset. Due to differences existent in the type of molecular descriptors on which a model can be based, non-parametric approaches are preferable. Moreover, probability-based methods are able to solve the problem of a region comprised in the convex hull but poorly described by the training set.

The mathematical complexity of the probability density function and distribution depends on the number of molecular descriptors of the (Q)SAR model. This complexity increases too much if the dimensionality is above 3, however some assumptions can be

made to overcome this limit [76] and algorithms have been developed in recent years to handle multivariate density estimation [77,78]. Furthermore, the "joint applicability domain" concept has been developed to consider the joint distribution of model's molecular descriptors and responses; this allows including the comparison between predicted and experimental values in the AD definition.

10.4.6 Structural Descriptors and Mechanism of Action: AD in local models for (bio)chemical activities

Molecules' potential (bio)chemical activities (such as toxicity) are usually determined by a particular part of the molecule itself rather than the entire structure. This complicates the definition of the model AD, since it is difficult to ensure that the needed structural features are present both in the training and the test sets. Considering the fact that the training set can only sample a small fraction of the reality, it is highly possible that the model would have to predict chemicals with unknown structural features; the prediction for this type of molecules is generally less reliable, this means that the model has to advise the user that an unknown fragment has been found in the tested molecule (the molecule is out of AD).

It is also possible that different sub structural fragments (also called biophores) are responsible for the same (bio)chemical activity. This structure-dependent AD definition can be used to group chemicals with similar biophores and these groups can then be used to build local (Q)SAR models. This type of approach results in a hybrid model which:

- addresses molecules to the appropriate model on the basis of the biophore found;

- warns user if an unknown fragment is present in the molecule.

Other parameters can be used to test if a molecule is inside or outside the AD, for example using structure similarity index such as the Tanimoto score [79], which is a widely used method to evaluate the similarity between chemicals. This score is defined by Rogers and Tanimoto as a "similarity ratio" given over bitmaps, where each bit represents the presence or absence of a characteristic. Considering for example two molecules (samples), the score is calculated dividing the number of common bits by the total number of non-zero bits in either samples.

Schultz et al. applied these kinds of approaches to classify chemicals for possible aquatic toxicity in six "modes" of toxic action (MOA): non-polar narcosis, polar narcosis, ester narcosis, soft-electrophilicity, pro-electrophilicity and respiratory uncoupling [80]. Test molecules have been predicted by different QSAR models depending on their MOA increasing the global performances of the model.

10.4.7 A classification model for Applicability Domain: the AD Metric

In 2009 Dragos at al. introduced the concept of *"...tool trying to find attributes that discriminate between compounds were well predicted and respectively mispredicted by a model..."* [81]. In this work, they treated the problem of assessing if a molecule is inside or outside the model AD as a classification model, introducing two elements:

- an AD metric, or "mistrust score", which is a function of the descriptors of the test molecule and of those of the training set molecules defining the AD;
- an unpredictability threshold denoting the highest acceptable mistrust score at which the misprediction risk is still acceptable.

The mistrust score has been defined as a sum of different contributors which will be introduced in the following paragraphs; most of these contributions can be calculated for the entire set of calculated molecular descriptors, in this case they are defined as model-independent contributors (I) or only considering descriptors which are part of the model (model-dependent contributors, D).

The Correlation Breakdown Count

When creating a (Q)SAR model, one of the first steps is the reduction of the number of molecular descriptors, in order to remove those with constant or near-constant value through the training set and to reduce correlated descriptor by keeping only one of them. Regarding the correlated descriptors, it is possible that this correlation only exists in the training set. The correlation breakdown count is a parameter which counts the number of correlations which are no more valid for a test molecule; in fact, it is expected that the probability of correctly predicting a novel chemical decreases with the number of descriptor pairs that fail to respect the correlation observed in the training set. This parameter can be calculated both as model dependent or independent.

The SQS Consensus Prediction Variance

As already observed for the previous parameter, correlated descriptors are reduced during classical (Q)SAR modelling. The Stochastic (Q)SAR Samples (SQS), on the contrary, is aimed at enumerating and building the maximum of possible (Q)SAR equations, alternatively using the other excluded correlated descriptors. Since the correlation between the descriptors could be no more valid for chemicals very different from those composing the training set, the predictions of a test molecule obtained by the models

generated by SQS may be different. The variance of the prediction may therefore give an indication about the chemical being inside or outside the AD. This parameter is obviously model dependent.

The out-of-Bound Descriptor Value Count

This parameter is based on a classical assumption for (Q)SAR AD definition: the reliability of the prediction decreases if the value of one or more descriptors calculated for the novel chemical to predict, lay outside the box defined by the training set compounds. The upper and lower limits for each descriptor have to be defined. The upper limit can be the maximum value of the descriptor in the training set, or a new value calculated as the sum of the mean value plus twice the variance value (if this is lower than the maximum value). For the lower limit the definition is quite similar, the highest between the minimum value of the descriptor in the training set and a new value obtained subtracting the variance from the mean is taken. The out-of-Bound Descriptor Value Count can be calculated both as model dependent and independent.

Dissimilarity-Related AD Metrics

Another classical assumption used for the definition of the AD metrics is that the prediction reliability decreases for molecules which are very dissimilar from those composing the training set. Three type of dissimilarity measures can be used:

- the **average of Dice dissimilarities** between the test molecules and all the molecules composing the training set; this approach uses z-normalized (average/variance-rescaled) descriptor values;

- the **biased version** of the previous approach uses weighted distance, assigning higher weight when the molecule in the training set is more similar to the tested one, these molecules are called nearest neighbours;
- an even **stronger bias** in favour of near neighbours can be introduced by considering only the nearest neighbours to the tested molecule in the distance calculation.

Dissimilarity-related approaches can be used both in model dependent and independent forms.

Combining different approaches: consensus AD Metrics

The most intuitive way to combine different approaches is to label chemicals as inside the AD of a model, only if their mistrust score is under the unpredictability threshold for all the calculated AD Metrics.

Dragos at al. however suggested a "fuzzy" AD definition which can be used not for an absolute inside/outside classification, but for the creation of a relative prioritization in terms of prediction trustworthiness. Since different AD Metric potentially cover different value ranges, it is important, prior to creating the consensus, to normalize the values (e.g., between 0 and 1). The mistrust score is replaced by the fraction of molecules with mistrust score higher than that calculated for the training set. These fractions, calculated for each AD Metric, are then summed to obtain the combined unpredictability value.

10.4.8 The VEGA approach for the applicability domain determination

VEGA [82] is a software platform developed as a "container" for (Q)SAR models, specifically focusing on those relevant within the REACH regulation. To produce reports that could be accepted as part of the dossier for REACH registration, a standardized applicability domain module has been implemented. The VEGA approach is mainly based on a comparison between the target chemicals and the most similar ones present in the dataset of the model. The similarity [83] is calculated combining a fingerprints-based distance calculation [84] with the distance calculated using molecular descriptors related to constitutional features, with the aim of integrating the information brought by the fingerprint with other structural information that is not encoded by the fingerprint itself, in order to obtain a more precise measurement of similarity. Three classes of molecular descriptors are calculated besides the fingerprints:

- basic constitutional features (number and type of atoms, number and type of bonds),
- ring features (number and size of rings, number of aromatic rings),
- functional groups (number of a set of functional groups, like esters, amines, etc.).

The Tanimoto [79] index is then used to calculate the distance between fingerprints whereas the cosine distance is used for the molecular descriptors. Finally, these four contributions are combined with a utility function: the four values are multiplied by a fixed weight then summed. The final similarity index has values in the range between 0 and 1, the latter meaning maximum similarity.

Once the similar compounds are identified, VEGA computes the so-called "applicability domain index" (ADI), which is calculated from the contribution of several parameters. Even if developed as an unified approach, the definition of AD of different type of models can be based on different parameters, for example depending on the type of descriptors on which the model is based on (fragments, physicochemical properties, topological information, etc.) or on the type of model (classification or regression). Considering for example the CAESAR model for Bioconcentration Factor (BCF), the ADI calculation considers the following parameters:

- **Similar molecules with known experimental values**: VEGA calculates an average similarity between the target molecules and the most similar ones found in the model's training set;
- **Accuracy (average error) of prediction for similar molecules**: is an evaluation of the model performance of the most similar compounds;
- **Concordance with similar molecules**: is the average difference between the target compound prediction and the experimental values of the similar compounds;
- **Maximum error of prediction among similar molecules**: is the maximum error among the most similar molecules;
- **Atom-centered Fragments (ACF) similarity check**: VEGA checks how the ACFs in the target molecule differ from those present in the training set;
- **Descriptor noise sensitivity analysis**: is an evaluation of the stability of the predictions with respect to small random changes of the descriptor values;

- **Model descriptors range check**: VEGA checks if the values of the molecular descriptors calculated for the target molecule are within their range calculated from the chemicals in the training set.

The AD approach integrated within VEGA is of particular interest. Besides considering the contribution of the structural information of the target molecule, in comparison with those of the model's training set, it also estimates how the endpoint and the chosen algorithm can influence the AD. As described above, by calculating the accuracy of prediction, the concordance and the maximum error of prediction, VEGA considers the model's capability to predict the specific endpoint, on molecules similar to the target one. On the other hand, the influence of the algorithm on the AD is considered by the analysis of the descriptor noise sensitivity.

11. Local models: when the AD is intrinsic in the data used to build the model

Depending on the composition of the training set, it is possible to distinguish between two types of (Q)SAR models: global and local. Global models are built from heterogeneous datasets and should in principle be able to predict any type of chemicals. On the other hand, local models are built using a subset of molecules, which share a specific set of features or properties that are related to the endpoint to predict [85].

The concept of local models is related to molecular similarity which, as already explained in the previous chapter, could be tricky do determine. However positive results in this direction have been obtained [4,86]. Another possibility is to build a

dataset of molecules that share features selected by human experts as relating to the biological effect.

To build a local (Q)SAR model, part of the available experimental data is discarded, this means that also part of the information about the structure-activity relationship is lost. This leads to model with a narrower predictivity of the chemical space. However, these local models should in theory be more sensitive to small structural differences, and show a higher accuracy with respect to global models. Several comparisons between local and global models have been made through years, each authors claiming the superiority of one approach or the other [4,85-93].

An interesting point while considering local (Q)SAR models is that their applicability domain is partially defined *a priori*, since the model should be used only with molecules that share the selected features with those in the training set. Moreover, as explained in the previous chapter, the applicability domain of models based on structural alerts (being either knowledge-based or statistically-derived) is also partially defined *a priori*. Considering for example the models for mutagenicity and carcinogenicity developed by Benigni et al. and implemented within the Toxtree software [14], each structural alert has been translated into a specific rule, which sometimes also includes exceptions. In this case the ruleset can be viewed as a set of local models, each model is applied only when its structural alert is present in the structure of the analysed compound. Therefore each structural fragment defines the prediction and provides *a priori* information of the applicability domain of the rule.

Chapter D.

Using structural features to improve the applicability domain definition: case studies

12. Applicability domain of knowledge-based models: a case study of Derek for Windows

In research published in 2011, Ellison et al. compared several methods to define the applicability domain of SAR models based on structural alerts [94]. One of the main aims of this study was related to how to consider the compounds which do not have any of the structural alerts defined within a knowledge-based model.

The authors started from the assumption that, if enough knowledge was available about a particular biological effect, the molecules which did not include any known structural alert could be predicted as inactive [95,96]. In this perspective, all the molecules not presenting any of the rules, should be inside the AD of the model.

The model selected for this study was Derek for Windows (DfW), a knowledge-based commercial system that includes SAR models for several endpoints. Five approaches for the determination of the applicability domain were compared: two types of fragment based approaches [97,98], the evaluation of the descriptor ranges, a structural similarity approach [99], and a fingerprint-based comparison [61]. The dataset chosen was the

publically available list of 4337 compounds and experimental Ames test results, collated by Kazius et al. in 2005 [100].

DfW evaluates the molecules on the basis of a set of rules, which includes both structural alerts and some molecular descriptors, and classifies the toxicity as certain, probable, plausible, doubted, improbable and impossible. If a molecules does not meet any of the rules, DfW's output is "nothing to report". To evaluate the impact of the AD methods, Ellison et al. decided to binarize the predictions, considering certain, probable and plausible outputs as positive (mutagenic). The other outputs (including "nothing to report") were considered negative (non-mutagenic). As for building (Q)SAR models, the determination of the AD of a model needs a training and a test phase. The original dataset was split in 10 combinations of training and test sets. For each combination, the test set included all the wrong predictions (false positives and false negatives) and a randomly chosen 10% of the correct predictions. Therefore, each AD method were trained and tested ten times, and the mean values were used to compare their capacity to improve the predictivity of the model.

The results of this study suggested that the approaches based on structural fragments could improve the AD definition of knowledge-based models. On the other hand, the results did not support the assumption that all the molecules for which the software could not make a prediction (labelled as "nothing to report") should fall within the AD, and be safely classified as non-toxic.

13. Atom-centered fragments to determine the applicability domain of (Q)SAR models

In 2009, Kühne et al. studied the possibility to use an approach based on the comparison of atom-centered fragments (ACF), to determine if a molecule falls within the applicability domain of a (Q)SAR model [101]. ACF are built starting from each atom of the molecule (in this study hydrogens were excluded), which is considered as the central atom. Each ACF is then defined through the atom-type and the number and type of bonding neighbours and associated bond types. Starting from the same central atom, several type of ACF can be built, depending on the number of atoms considered in each bonding direction. This particular type of approach, as stated by the authors, should be able to determine the AD *a priori*, making it model-independent. The authors based their approach on important evidences of the usefulness of ACF in improving the AD definition and (Q)SAR models' performance [95,96,102,103,104,105].

To study the ACF approach, the authors used three models as case studies (including both regression and classification models): a structural fragments-based model for the estimation of the logarithmic molar water solubility [106], a classical molecular descriptor-based model for the prediction of the logarithmic air/water partition coefficient [107], and a structural alert-based model to predictively discriminate between narcosis-level and excess-toxic compounds with respect to Daphnia [108]. For each model, the training set was available, and the authors also found considerably larger external dataset, to use as test set. Finally, four types of ACF were defined, with increasing complexity and also three types of matching mode were defined, with

increasing strictness. As expected, all the models performed worse on the external dataset then on their training set. Considering the four types of ACF and applying the matching modes to the external dataset, the performance of the three models improved. Obviously, the more specific an ACF is, the less number of external molecules are expected to fall within the AD. Similarly, the strictest matching mode, would exclude more molecules from the AD. The results of this study confirmed that atom-centered fragments can help to improve the definition of the applicability domain of both QSAR and SAR models. However, using very specific fragments and/or strict matching modes may lead to high predictivity, but on very small subset of molecules. For example, applying the strictest matching modes to compare the most complex ACFs, lead to R^2 values very close to those obtained on the training sets, but with an applicability domain reduced to a 1 to 3% of the original dataset. The most promising ACF-based method to improve the AD definition resulted that based on moderate complex ACF and using the less strict matching mode. Also in this case, the performance on the subset defined by the AD are very high and close enough to those on the training set. Anyhow, the number of compounds which fell within these ADs was also low, spanning from 11 to 13% of the original external datasets.

14. A chemical classes-based evaluation of the AD of (Q)SAR models: a case study from the ANTARES project

ANTARES [55] was a LIFE+ funded project aimed at reducing the gap of knowledge on which *in silico* methods were available, and could be used, to evaluate properties

relevant within the REACH regulation [109]. Among the main outcomes, the ANTARES project produced an on line list of commercial and freely available models for several endpoints. The predictive performance of these models were evaluated using external datasets, in order to provide a broad view of which software are able to give the most reliable predictions and under which conditions.

Within ANTARES, "Action 6" specifically aimed at identifying "...boundaries for best use of models (applicability domain) and of the assessment factors". The results of this action are reported in the project's official document *"Report with the discussion and identification of the applicability domain for each validated model"* (deliverable 26), which is available upon request. Two approaches have been reported, one is based on the study of the relationships between effectiveness of prediction and chemical classes, whereas the other relates the same effectiveness with mode of action.

To study the relationship between AD and chemical classes, the predictions obtained were clustered into two groups: molecules correctly predicted and molecules wrongly predicted. For endpoints characterized by continuous values (e.g. BCF, LD50, etc.), the experimental variability of bioconcentration factor, determined by Dimitrov et al. [15] was used as threshold. Molecules whose predicted activity differed more than 0.7 log unit from the experimental value were clustered as wrongly predicted. The occurrence of different functional groups (FGs) within each dataset was calculated using the ISSFUNC module included in the Toxtree software [14]. The distribution of functional groups was compared between the two clusters of molecules and, using a frequency threshold, the chemical classes were included or excluded from the applicability domain.

Finally, using a Canonical Discriminant Analysis (CDA) the results from (Q)SAR predictivity and ISSFUNC were combined, obtaining correlation coefficients for each classes. Higher values of this coefficient means that the molecules belonging to that class are predicted very differently (better or worse) than the whole dataset.

The results showed that using chemical classes can improve the definition of (Q)SARs' AD. However, this is not always true. Indeed, the method used seems not able to improve the AD definition of models whose performance are good on the whole dataset used for the study. More interesting results were obtained on less reliable models.

Part 2 - Work done and results: Using structural features to improve the applicability domain definition of (Q)SAR models

Chapter E.
Materials and methods

15. Research framework: the LIFE+ project ANTARES

15.1 The main aim of ANTARES: assessing the performances of available (Q)SAR models

The LIFE Programme is the European instrument which aims at supporting environmental and nature conservation projects throughout the EU. In particular, within the LIFE+ Environment Policy and Governance component, a particular attention is paid on projects that offer significant environmental benefits, for example process or efficiency improvements. In the period from 2007 to 2013, LIFE+ also funded projects that improve the implementation of EU environmental legislation, that build the environmental policy knowledge base, and that develop environmental information sources through monitoring. On January 1st 2010, LIFE+ started to fund a three years project called ANTARES [55], which dealt with the assessment of non-testing methods (NTM), including (Q)SAR models, for their possible use within the REACH regulation framework [109].

The main aim of the ANTARES project was to collect information about *in silico* predictive models developed so far, which could be useful to obtain the (eco)toxicological information required by REACH. In particular the objective were:

- to verify the possible use and performance of the non-testing methods for REACH;
- to identify requirements and constraints originating from the REACH legislation which may affect the non-testing methods;
- to identify safety assessment factors for the non-testing methods;
- to identify the best applicability criteria for a safer use of the non-testing methods;
- to integrate different non-testing methods, achieving superior performance;
- to disseminate the results;
- to promote non-testing methods for legislative purposes.

Within ANTARES, the activities not only regard a simple compilation of what's available. The so-called "Action 5" and "Action 6" dealt with the validation of non-testing methods and the identification of their applicability domain. Moreover, "Action 7" dealt with the integration of different models for the same endpoint, to improve the overall performance and coverage of applicability. For each endpoint described within the REACH legislation, a list of available models were built. ANTARES also searched for available experimental data, which were necessary for the models' assessment. Eight endpoints were selected on the basis of the availability of models and experimental information, leading to the assessment of more than 50 (Q)SAR models [56,57,58,59,60].

15.2 The method used within ANTARES

15.2.1 Performance of regression and classification models

Using experimental data from reliable sources, a dataset was built for each endpoint. In case of multiple values for the same chemical, the arithmetic mean was calculated. Each structure present was double-checked using online databases such as ChemSpider [110,111] and ChemID*plus* [112,113], verifying the correct match of chemical names, CAS registry numbers and SMILES. The inorganic compounds, mixtures of either compounds or isomers, and molecules with insufficient information were removed. Finally, the salts were converted to their acidic forms (by removing the counterions).

The datasets were predicted using the available (Q)SAR models and their performance was evaluated. For regression models two statistical parameters were considered:

- The coefficient of determination (R^2), which determines how closely a linear function fits a set of data.
- The root-mean-square error (RMSE), which is a measure of the accuracy based on the differences between values predicted by a model and the values actually observed.

Classification models were evaluated using the confusion matrix, a 2x2 matrix containing the four possible outcomes of a classifier:

		Predicted Value	
		Toxic	Non-toxic
Experimental Value	Toxic	TP	FN
	Non-toxic	FP	TN

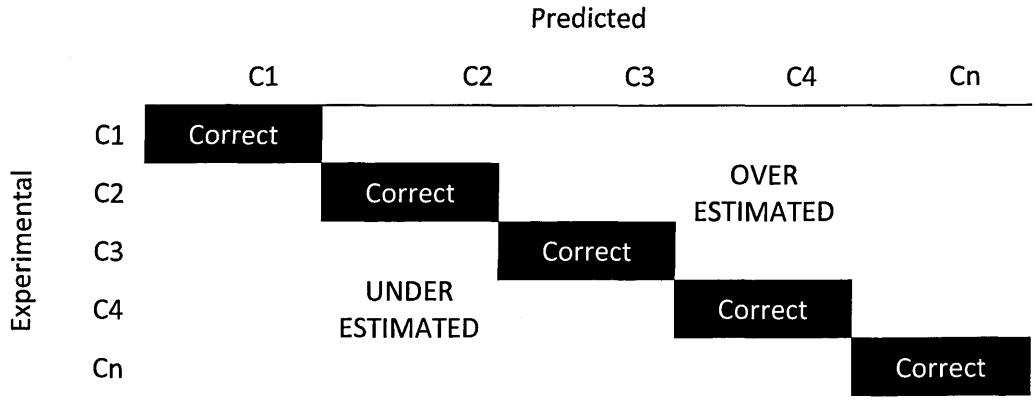
Where:

- The True Positive (TP) class includes the compounds correctly predicted as toxic;
- The True Negative (TN) class includes the compounds correctly predicted as non-toxic;
- The False Positive (FP) class includes non-toxic compounds predicted as toxic;
- The False Negative (FN) class includes toxic compounds predicted as non-toxic.

From the confusion matrix, three parameters were calculated to compare the models' performance:

- Accuracy ((TP+TN)/Total): the capacity of the model to correctly predict a molecule
- Sensitivity (TP/(TP+FN)): the capacity of the model to correctly predict a toxic molecule
- Specificity ((TN/(TN+FP)): the capacity of the model to correctly predict a non-toxic molecule

Within several regulatory frameworks, thresholds of concern have been introduced for continuous endpoints. These thresholds were adopted within ANTARES to evaluate the capacity of QSAR models to correctly "classify" the molecules for regulatory purposes. More details about specific endpoint-related thresholds will be given in the next sections. For endpoints such as BCF and acute toxicity, multiple thresholds have been defined, leading to the formation of more than two classes. In this case it was not possible to define a prediction as TP, TN, FP or FN, therefore the confusion matrix were adapted and simplified.



In the example above, C1 to Cn are the classes defined by the thresholds adopted within a specific regulation. All the molecules laying on the main diagonal of the matrix were correctly predicted by the model. The meaning of the Accuracy parameter was therefore extended to include all these molecules, and calculated as the ratio between the number of correct predictions and the total number of molecules predicted.

15.2.2 Considering the applicability domain and the models' training set

Considering the whole dataset for the assessment of (Q)SAR models only gave partial information about the performance. As explained in Part I – Chapter A, predictive models are generally better while evaluating molecules used to train them. The assessment could be biased by a high overlap of the dataset used to evaluate the model and that used to build it, leading to an over-estimation of the performance. To consider this and therefore to assess how the models behave on "new chemicals", for each model the dataset were split in two classes: molecules present in the model training set and those not present in it. The performance on these two classes was then evaluated.

Whereas the presence of known chemicals could lead to an over-estimation of the performance, molecules outside the applicability domain could decrease it. The applicability domain information provided with the models' prediction was used to classify the molecules as "within the AD" and "outside the AD". Again, the models were evaluated on these two classes separately.

Ideally, to evaluate molecules whose experimental values are not available, more than one (Q)SAR model should be used. Moreover, these models should be provided with information about their applicability domain, so that users can check if the molecule to predict falls within it. Within this perspective, the classifications described above were merged to assess the models performance on new chemicals which fall within their AD, further checking the usefulness of AD on chemicals for which the models do not have any information.

16. Case studies considered

16.1 The endpoints

Three endpoints were chosen among those considered within ANTARES: the bioconcentration factor (BCF), mutagenicity and rat oral acute toxicity. In order to get the most comprehensive view of the application of AD to (Q)SARs, the selected endpoints were characterized by different types of values (e.g. continuous or discrete) and different algorithms used to build (Q)SAR models. Moreover, depending on how chemicals interact with the organisms and target the biological effects, certain endpoints (such as BCF) are usually easier to model. On the other hand, modelling

endpoints such as the acute toxicity could be more difficult, because chemicals may target different types of biological processes, which lead to the same effect. In this case models could be more strictly related to their dataset.

16.1.1 Bioconcentration factor

Bioconcentration is a process that results in an organism having a higher concentration of a substance than is in its surround media (e.g. stream water). The bioconcentration of a substance is related to its octanol/water partition coefficient (K_{ow}). Bioconcentration factor (BCF) is the concentration of a particular chemical in a tissue per concentration of chemical in water, and is expressed as L/kg. This property characterizes the accumulation of pollutants through chemical partitioning from the aqueous phase into an organic phase. Aquatic organisms may accumulate chemical substances either directly from the environment, or through the food chain. Toxic chemicals with a high bioaccumulation potential may represent a dangerous threat both for animals and humans. BCF is commonly used as a first indicator for bioaccumulation. The most common experimental method to estimate the BCF is the "flow-through fish test" [114] whose guidelines have been defined within the Organisation for Economic Co-operation and Development (OECD) test guidelines 305 [115]. These kinds of tests for BCF are time-consuming and expensive, and are also characterized by a great variability [116]. For these reasons, (Q)SAR models have become more and more important in the evaluation of the bioconcentration factor.

Since BCF is expressed using continuous values, it is modelled using regression approaches. The octanol/water partition coefficient can be commonly found among the molecular descriptors of QSAR models for BCF. However, several physicochemical properties have been identified, which influences the relationship between K_{ow} and BCF [114]. For example, the molecular weight seems to play a key negative role in the bioaccumulation of chemicals. Above a certain dimension, it seems that the steric hindrance prevent the molecules to pass through the membranes. A cut-off limit of 700 for the molecular weight has been generally accepted [117]. Other parameters which may affect the bioconcentration, are the lipid solubility, biodegradability, volatility and the metabolism of the organism.

BCF is expressed as a continuous value and therefore modelled using regression approaches. However, BCF is considered within several regulatory frameworks. For each regulation one or more thresholds have been set:

- LogBCF > 2 as established for the Chemical Safety Assessment (CSA)
- LogBCF > 2.7 as established for the Classification and Labelling (C&L)
- LogBCF > 3.3 as established for the Persistent / Bioaccumulative / Toxic classification (PBT)
- LogBCF > 3.7 as established for the very Persistent / very Bioaccumulative classification (vPvB)

Considering these thresholds, (Q)SAR models can be also evaluated as classifiers.

16.1.2 Oral rat acute toxicity

As defined within Annex I of the REACH regulation:

> *Acute toxicity means those adverse effects occurring following oral or dermal administration of a single dose of a substance or a mixture, or multiple doses given within 24 hours, or an inhalation exposure of 4 hours.*

Acute toxicity used to be assessed with the "median lethal dose" *(LD_{50})* approach, which indicates the dose that kills 50% of animals tested within 24 hours [118]. For a classical LD50 study, laboratory mice and rats (of both sexes) are species typically selected. In 1987, OECD adopted the Test Guideline 401 for acute oral toxicity testing, using LD50. This guideline was subsequently removed and the LD50 test requirement was abolished. Nevertheless, the available experimental LD50 values can still be used to develop (Q)SAR models.

From the point of view of QSAR modelling, the situation is analogous to that described for BCF. LD_{50} is represented by continuous values and the models can be evaluated using the coefficient of determination (R^2). Within the Annex I, Part 3.1.3, of the CLP European regulation, four toxicity categories have been defined using LD_{50} thresholds, called "Acute Toxicity Estimate" (ATE). These categories (Table 1) were used to evaluate the models in classification, as for bioconcentration factor.

Table 1. Categories of Acute Toxicity Estimate (ATE) identified by the CLP European regulation

Categories	ATE limits (mg/kg)
Category 1 (C1)	ATE ≤ 5
Category 2 (C2)	5 < ATE ≤ 50
Category 3 (C3)	50 < ATE ≤ 300
Category 4 (C4)	300 < ATE ≤ 2000
Not classified (NC)	ATE > 2000

16.1.3 Mutagenicity

Mutagenicity can be defined as "the ability to cause permanent mutation in DNA sequence" and is a critical component of carcinogenesis.

A common and accepted method to obtain experimental values for mutagenicity is an *in vitro* approach called Ames test [119,120]. This test uses several strains of the bacterium *Salmonella typhimurium* that carry mutations in genes involved in histidine synthesis. The method tests the capability of a chemical in altering the DNA in a way that the mutated genes are reverted to their functional form, therefore restoring the possibility for the cell to survive and grow in a histidine free medium.

Typically, the Ames test is used to classify the substance, which is labelled as mutagen or not mutagen. Therefore, *in silico* models for the prediction of this endpoint are classifiers and their performance can be evaluated using parameters such as accuracy, sensitivity and specificity.

16.2 Dataset used

16.2.1 Bioconcentration factor

To test the available models, a dataset of compounds with known structure and high quality data on experimental BCF has been used. This dataset has been assembled using five datasets:

- Dimitrov et al., 2005 [15] (511 compounds),
- Fu et al., 2009 [121] (138 compounds),
- Footprint PPDB [122] (159 compounds),
- CEFIC LRI [123] (551 compounds),
- Arnot & Gobas, 2006 [124] (759 compounds).

The final dataset was composed of 860 compounds. For compounds present in more than one dataset and/or with more than one experimental BCF value, the mean were calculated and used.

16.2.2 Oral rat acute toxicity

The dataset used for acute toxicity was composed of 7417 organic compounds. Structures and experimental values were obtained from the dataset used by US Environmental Protection Agency (EPA) to develop the (Q)SAR model for acute toxicity integrated within the T.E.S.T. [125] software. EPA's dataset originally contained 7420 molecules, however three were removed due to problems of compatibility with the software used. The acute toxicity values in this dataset were derived from LD50 tests on rats via oral administration, and were expressed as -Log(mol/kg_{bw}), where mol is the

dose administered expressed in mols, and kg_{bw} is the weight of the rat. As described within the results obtained by the ANTARES project on acute toxicity [60], all the (Q)SAR models tested uses LD50 expressed in mmol/kg, so the experimental values were converted accordingly. Moreover, the CLP regulation requires LD50 expressed as mg/kg, so the experimental values were also converted to meet this requirement. Both values were used within ANTARES to assess the models, showing that the performance obtained using the values in mmol/kg were better than those obtained with mg/kg.

16.2.3 Mutagenicity

The dataset used for mutagenicity was composed of 6065 molecules, and included both structural information and experimental values from Ames test. The data was obtained from a dataset compiled by Hansen et al. in 2009 [126], which contains 6512 chemicals obtained from different sources:

- Chemical Carcinogenesis Research Information (CCRIS) [127] (2539 compounds)
- Kazius et al. 2005 [100] (2224 compounds)
- Helma et al. 2004 [128] (138 compounds)
- Feng et al. 2003 [129] (391 compounds)
- Virtual International Toxicology Information Centre (VITIC) [130] (1194 compounds)
- Genetic Toxicology Data Bank (GENE-TOX) [131] (26 compounds)

The dataset compiled by Hansen et al. contains the canonical SMILES representation of the structures, the Ames test results (mutagen or non-mutagen), and the references. The dataset were checked and cleaned for the presence of duplicates, salts, mixtures, and ambiguous compounds, resulting in the final set of 6065 molecules.

16.3 (Q)SAR models selected

Several type of modelling algorithm can be used to develop SAR and QSAR models for different endpoints, for example on the basis of the type of values used to represent them (continuous or discrete). Besides considering the endpoint-related difference of (Q)SAR models, as introduced in the previous paragraphs, this study also aimed at assessing how a chemical classes-based approach for the AD definition, could improve the performance of different type of (Q)SAR models. In particular, this study covered commercial and freely available software, regression and classification models, as well as models based on molecular descriptors and structural features (e.g. structural alerts). Table 2 shows all the software assessed. In order to maximize the exploitation of the results obtained from this study, the models considered should be easy to integrate with the AD information obtained. An even better possibility would be to automate this integration.

Table 2. Software considered for the study of the application of a chemical classes-based AD.

Software	Referred as (within this work)	Endpoint	Algorithm	Output type	AD Info provided	Training set available
ACD/Labs ToxSuite v2.95 [132]	ACD	LD50	Expert Knowledge and Classification Structure-Activity Relationship	Continuous	Yes	Yes
Simulation Plus ADMET Predictor v6.0 [20]	ADMET Predictor	LD50	Artificial Neural Network Ensemble	Continuous	Yes	Yes
TerraBase Inc. TerraQSAR v1.2 [133]	TerraQSAR	LD50	Probabilistic Neural Network	Continuous	No	Yes
U.S. EPA Toxicity Estimation Software Tool (T.E.S.T.) v4.0.1 [134]	T.E.S.T.	LD50	Five models: four (Q)SARs and a consensus	Continuous	Yes	Yes
Accelrys TOPKAT (Discovery Studio 3.1) [135]	TOPKAT	LD50	Combination of 19 chemical classes-related QSAR models	Continuous	Yes	Yes
CAESAR 2.1.13 (VEGA 1.0.8) [136]	CAESAR for BCF	BCF	Neural network based on 8 molecular descriptors	Continuous	Yes	Yes
CAESAR 2.1.12 (VEGA 1.0.8) [137]	CAESAR for Mutagenicity	Mutagenicity	(Q)SAR statistical model based on Neural Network + rule based model	Classes	Yes	Yes
SARpy 1.0.6-dev (VEGA 1.0.8) [22,59]	SARpy	Mutagenicity	Statistically relevant structural features	Classes	Yes	Yes
Benigni-Bossa Ruleset 1.0.0-dev (VEGA 1.0.8) [14]	Toxtree	Mutagenicity	Knowledge based structural alerts	Classes	Yes	Yes

The VEGA platform [82] is a free and open source software developed within our research group, with the aim of providing (Q)SAR models that can be used within current regulatory frameworks (e.g. REACH). The idea was to use models present within the VEGA platform as case studies, giving the possibility to integrate the applicability domain tools already present, with the findings obtained. Two models were initially available for the endpoints chosen: the CAESAR models for bioconcentration factor [136] and mutagenicity [137]. Both these models were interesting since they were based on different approaches. The BCF model was a classical molecular descriptor-based QSAR, whereas that for mutagenicity was a three-step hybrid model (a descriptor-based SAR module, followed by two consecutive structural alerts modules). Subsequently, two more models for mutagenicity were included within VEGA, and were considered for this study because their approaches differed from the CAESAR one, providing additional points of view for the AD study. The VEGA model developed using the SARpy software [22,59], which will be described later, was considered because it was based on statistically-derived structural fragments. The Benigni-Bossa rulebase for mutagenicity [14] was also integrated within VEGA, and considered for this study as a case study for knowledge-based models.

No models for acute toxicity were present within VEGA, and none have been integrated so far. However, this endpoint was considered interesting due to its complexity and to the availability of (Q)SAR models and experimental data. The results of the ANTARES assessment on five software for oral rat LD50 prediction were used as starting point in this study. T.E.S.T. [134] is the only freely available software considered,

and provide also the training and test sets the developers used to build and validate it. This software combines the predictions of four QSAR models developed using different approaches, considering also their AD, to obtain a consensus prediction.

The other software do not provide information about the training set used. The developers of ACD and TerraQSAR provided us the training set and we compared them with ours using two software: PerkinElmer ChemFinder [138] and Chemaxon InstantJChem [139]. TerraQSAR's developers provided us with the list of chemicals in common between our dataset and the training set. TOPKAT gives the possibility to export a list of the most similar compounds in the model dataset for each chemical evaluated (up to five). Each molecule is provided with the experimental value and a similarity measure; molecules with similarity index of 1 (100%) were manually checked to verify which of them were in common between the model's training set and our dataset.

The intrinsic complexity of acute toxicity is reflected by that of the available models. Both ADMET predictor and TerraQSAR were based on neural network, whereas the other software were built using multiple models (19 different models in the case of TOPKAT).

16.3.1 Applicability Domain determination approaches integrated in the selected models

All the software considered in this study, with the exception of TerraQSAR, provide information about the reliability of the prediction. Table 3 gives an overview of which methods have been included by the developers. The most common approach found in

the software considered in this study is the check of the descriptor range. TOPKAT includes two types of this approach: the commonly used and simple comparison of the molecular descriptor values between predicted molecule and the training set, and the Optimal Prediction Space check (both methods are explained in Chapter A). The comparison between structural fragments of the target molecule and those obtained from the training set (or a subset, for example the most similar compounds) is the second most common approach adopted. Similarity is also a parameter considered by three out of five software, both VEGA and ACD calculate the structure similarity, whereas T.E.S.T. utilized the molecular descriptor values.

Table 3. Methods for the assessment of AD included in the software considered in this study.

		VEGA **	ACD	ADMET Predictor	TOPKAT	T.E.S.T.
Methods based on similarity between target molecule and model training set	**Presence of structurally similar compounds**	X	X			
	Distance-based method (calculated with descriptors)					X
	Fragments comparison	X			X (1)*	X
Methods including experimental and predicted values of training set molecules	**Prediction-Experimental Concordance**	X	X			X
	Accuracy of prediction for similar molecules	X				
	Maximum error of prediction among similar molecules	X				
	consistency of experimental values for similar molecules		X			
Methods based on descriptors range	**Model descriptors range check**	X		X	X (2)*	X
	Accelrys Optimal Prediction Space (OPS)				X (3)*	
Other methods	**Descriptors noise sensitivity analysis**	X				

* TOPKAT performs a three-stage analysis, from step 1 to step 3.

** VEGA includes CASEAR for mutagenicity and BCF, Toxtree, and SARpy. The applicability domain is calculated using the same parameters, with some exceptions depending on the type of descriptors and output (e.g. continuous or discrete values)

The endpoint and the ability of a model to obtain reliable prediction are also considered by most of the models, even if in different ways. VEGA includes three parameters for keeping into account these aspects; the concordance between the predicted value obtained for the target molecule and the experimental values for the most similar compounds present in its dataset; the accuracy and maximum error obtained while predicting the similar compounds. Also T.E.S.T. and ACD compare the prediction for the target molecules with the experimental values of molecules in their dataset. Furthermore, ACD also considers if the endpoint has similar values for molecules similar to the target one. Finally, the algorithm strength is also considered by VEGA: introducing small changes in the molecular descriptors, the software checks how the prediction is affected.

The software also differ on how these approaches are applied and the output. VEGA combines all the obtained values to calculate an "Applicability Domain Index" (ADI), which spans from 0 to 1. ADI is then used to label the prediction as reliable or not. ACD also uses the parameters to compute a Reliability Index (RI), again spanning from 0 to 1, which is provided with the prediction. The user has to decide a RI threshold to use to consider a prediction as reliable. Generally, during validation by the developers, predictions providing a RI less than 0.3 were considered unreliable. TOPKAT adopts a three-stage analysis, at each step the molecules that do not match the constraints are excluded and labelled as out of the applicability domain. T.E.S.T., as already explained, predicts the molecules using multiple models, for each of them perform an applicability domain check, and the molecules falling out of it, are not predicted.

16.4 Software used

16.4.1 Standardization of the molecular representation

As introduced in chapter A, many (Q)SAR modes and chemoinformatic tools utilize the Simplified Molecular-Input Line-Entry System (SMILES) [25] formalization to read the molecules in input. SMILES are usually computed using specific software, which however produce different formalization for the same structure, depending on the algorithm implemented. Two SMILES type were used in this study, the VEGA formalization and the canonical SMILES [25].

With the aim of providing useful tools for (Q)SARs users and developers, several freely-available chemoinformatics tools have been developed in the last years in collaboration with Kode S.r.l. [140]. These software are based on the same libraries developed for VEGA, using the Chemistry Development Kit (CDK) [141] as infrastructure. The istMolBase software is an easy-to-use tool for dataset visualization and management; molecules can be imported in both SMILES formats and Molecular Design Limited (MDL) sdf files [26], and exported as standardized SMILES. istMolBase also includes two interesting features: it is able to neutralize the molecules and it can perform the SMARTS [142] matching. This tool were used to convert all the dataset used in the VEGA SMILES format.

Open Babel [143] is an open source toolbox including several tools useful in chemoinformatic. Among its feature, Open Babel is able to convert between most of commonly used formats for molecular representation, including the so-called "canonical SMILES". As suggested by the name, this formalization aims at providing an unequivocal

way to convert the molecular structure (including properties such as the stereochemistry) in SMILES format [144]. Open Babel was used to obtain the Canonical SMILES representation of the dataset used within this study.

There are two main reasons behind the use of two type of standardized SMILES formats: comparing how much the formalization could influence the results obtained and using a formalization that could be easily implemented within our tools. This second reason applies to the VEGA format.

16.4.2 Structural feature extraction and validation

istChemFeat is another software included within the aforementioned In-Silico Tools. It was developed for dataset analysis, with the aim of searching for relevant chemical features. The software includes a list of more than 300 functional groups and atom-centered fragments. istChemFeat requires a dataset where each molecule is assigned to a class. The application in turn produces a list of the main chemical features with their number and percentage in each class.

istChemFeat leave the analysis of the results to the user. A more advanced tool, called istRex, were developed on the basis of the approaches used within this work. istRex, which is still in beta version, takes in input a list of SMILES associated with the binary property and derives rules from the extracted structural feature. As for istChemFeat, the molecules are checked against an internal library of structural features (including functional groups, atom-centered fragments, relative position of different atom-types in a ring, etc.), and for each feature the software statistically analyses

whether its presence is able to improve the discrimination between the two classes defined by the target property. istRex utilizes the *p*-value to select which features are significant for the classification. Furthermore, using the same approach, istRex can analyse each subset of molecules to extract secondary rules. For example, if the aim is to extract structural features related to a particular toxic event, istRex analyses all the features present in the dataset and extracts only those whose ratio of toxic molecules is significantly higher than that calculated on the entire dataset. Each primary feature is then used to extract subset of molecules and the software extract secondary rules which either significantly increase or decrease the ratio with respect to the primary rule. The user can also decide how many level of sub-rules istRex should try to reach.

While both istRex and istChemFeat are based on a library of predefined structural features, the freely available software SARpy [22] builds the library from the provided dataset and identifies structural fragments statistically related to the chosen property. The current version of the software only works with discrete binary classes (e.g. toxic / non-toxic). Several parameters can be set which can influence the fragments extracted, such as the minimum and maximum number of atom in the fragment and the minimum number of its occurrence. Moreover, SARpy gives the possibility to evaluate the two classes together or separately and lets user decide how to optimize the results (maximize predicted rate, minimize errors, etc.).

Could SARpy be used to identify fragments, either scaffolds or small groups, related to the model's ability to obtain reliable predictions? In other words, the idea was to try to identify structural features which can be used to warn the user if a molecule is outside

the AD of a model or, on the other hand, to identify fragments which can be used to label a prediction as reliable.

17. Approaches to study the applicability domain of (Q)SAR models

Starting from the promising results obtained so far, while studying the applicability domain of (Q)SAR models as a function of sub-structural features such as atom-centered fragments (ACF) and structural alerts (SA), three different statistical approaches based on structural fragments have been investigated. A simplified and general approach considered the chemical composition and size of molecules, to identify outliers on the basis of uncommon characteristics. Starting from the concepts of statistically-derived atom-centered fragments [101] and of "modelling the error" [81], the correct and wrong predictions of models have been correlated with structural sub-features. In one approach, these features were statistically built and extracted from a dataset used as a training set. In a second approach, the dataset has been compared with a library of functional groups and ACFs, extracting the features which seem to be statistically related with the discrimination between wrong and correct predictions. To study the possibility of an *a priori* determination of (Q)SARs AD based on structural features, these approaches have been also applied on the endpoint values, studying whether the balance of these properties can affect the prediction capabilities of particular chemical classes.

17.1 Atomic composition and molecular size: studying a simplified and general approach to determine the AD

To interact with the biological macromolecules (DNA, proteins, etc.), chemicals have to be of the correct size and composition. Both ACF and SA include information about the atomic composition of molecules or a portion of them. The molecular weight is commonly found among the molecular descriptors calculated by chemoinformatic software, and it provides general indications about the molecular size. Atomic composition and molecular weight has been used as a simplified and general approach to study the AD of (Q)SAR models. The main idea was to identify outliers on the basis of both their chemical complexity, and their "borderline" characteristics. As explained in Part I, (Q)SAR models depends on the chemical similarity, therefore molecules with an uncommon composition or size could represent a problem for (Q)SAR models.

The method used to study the relationship between models' prediction performance and the selected properties (molecular weight and composition) was rather simple and was aimed at studying the AD of "simple" models, such as those developed for BCF.

The commercial software Discovery Studio 3.0 (DS3) by Accelrys Software Inc. [145] has been used to calculate the molecular weight and composition of the molecules. In particular, the software calculates both the empirical formula and the percentage (in weight) of each atom-type present in a molecule; this last parameter was used to classify the molecules of the dataset.

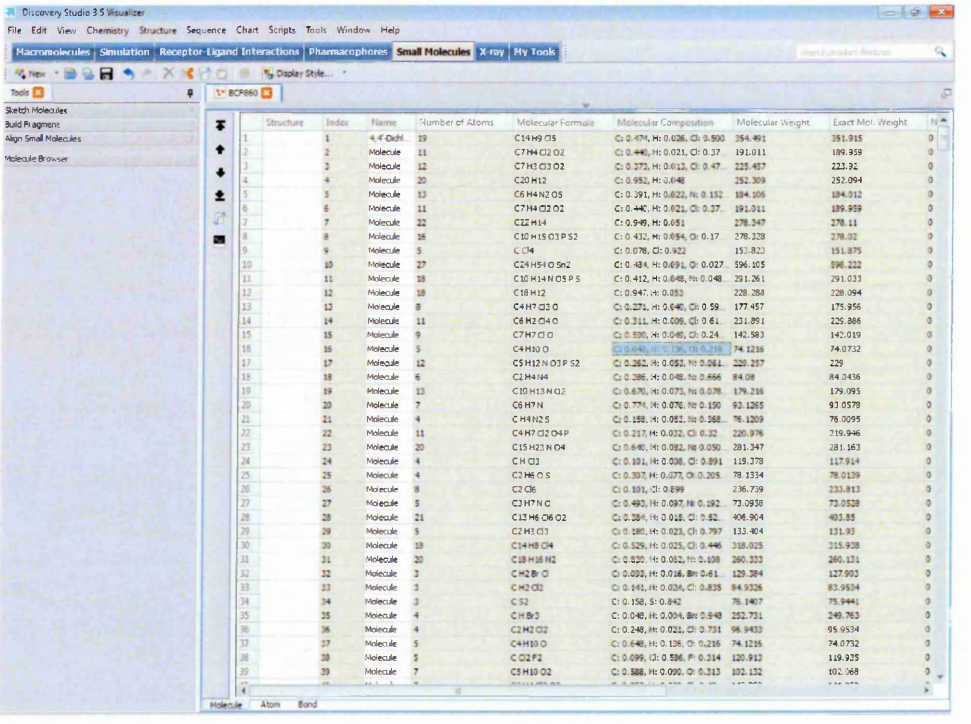


Figure 1. Accelrys Discovery Studio automatically calculates basic properties such as molecular weight, formula and composition.

The output generated by DS3 was imported in Microsoft Excel and the column containing the percentage composition was split in order to have each atom-type in a different column. For each row, the atom-types were aligned to have a consistent column-type association. Using these columns as well as that containing the molecular mass information, the dataset could be ordered by each atom-type. Moreover, the columns were combined to obtain the total percentage of a certain class of atoms (e.g. halogen atoms).

The main aim was to obtain thresholds for the selected properties, to be used for the determination of the applicability domain. These thresholds needed to be tested on an external set of molecules. For this reason the dataset was split in a training and a test

set. The molecules were sorted by their experimental LogBCF values in ascending order, one out of every four were assigned to the test set and the other to the training set. This approach was used to obtain a homogeneous sample of all the molecules in the dataset, based on their activity. Furthermore, a 10% manual leave-more-out approach was used to analyse the obtained training set. A random number was assigned to each molecule using the Excel function RAND(), the training set was sorted using these values and one in every 10 molecules was removed and assigned to a prediction set. This procedure was repeated five times.

A histogram-based approach was chosen to classify and represent the dataset. The training set was ordered by each chosen parameter and classified by applying thresholds. Three parameters were obtained for each class:

- the average error, calculated as the mean of the absolute difference between predicted and experimental values;
- the coefficient of determination (R^2) calculated between predicted and experimental values;
- the number of molecules.

These parameters were then graphically represented using a combinations of histograms. R^2 and average error had the same scale and were plotted together. Classes with higher R^2 were expected to have low average error, leading to opposite trends (if a trend between the analysed property and the model's performance existed). Another histogram was plotted, using a different scale, to show the distribution of the number of molecules. The distributions of the training set were analysed, searching for

significant correspondences between the increase or decrease of the chosen parameter (e.g. molecular weight) and the model's prediction performance. The results obtained with the five iterations of the leave-more-out approach were compared and the thresholds were then tested on the test set. Moreover, applying the ANTARES approach, these thresholds were also tested considering separately the molecules present in the model's training set (in-model-training) and the "new molecules" (out-model-training).

17.2 Generating structural fragments statistically-related to the models' predictivity: SARpy

The VEGA software classifies molecules as mutagen, non-mutagen or suspect mutagen. Since a conservative approach is generally preferable when predicting toxicity for regulatory purposes, the suspect mutagen molecules were considered as mutagens. The results of the comparison between experimental values and the two obtained classes predicted by VEGA were organised in the classic confusion matrix composed by TP, FP, TN and FN. The two type of errors were considered separately. This seemed reasonable, since the causes that lead a model to predict a toxic compound as non-toxic (FN) could be different from those related to the opposite error (FP).

The dataset was initially divided into a training and a prediction set. In order to have representative of the four type of predictions in both sets, 2/3 of the molecules for each class were used for training set and the remaining 1/3 were used for the prediction set.

To develop the SARpy models for TP/FP and TN/FN molecules, both training and prediction sets were split to have each subset containing only the molecules associated with the target labels. The training sets were then used with SARpy to extract the

structural features. The obtained rulesets were used to predict the reliability of the predictions obtained by the (Q)SAR model. Molecules predicted as TP or TN were considered as within the model applicability domain, whereas FP and FN were considered out of it. The performance of the (Q)SAR models were than analysed using an approach similar to those used within the ANTARES project. Molecules part of the SARpy training set were considered separately to those in the prediction set to test the reliability of the ruleset on "new" chemicals, which were not used for its development. For the same reason, the dataset was also split in two groups, one containing the molecules present within the (Q)SAR model training set and the other containing new molecules.

The reliability of the applicability domain defined using SARpy rules was compared to that included within VEGA. The performance of the (Q)SAR models were evaluated for molecules within and outside both applicability domain, also considering molecules in/out the training sets of both the SARpy ruleset and the(Q)SAR model.

17.3 Using a library of predefined structural features to compare the applicability domain of (Q)SAR models for the same endpoint

In: Gonella Diaza R et al. Comparison of in silico tools for evaluating rat oral acute toxicity. SAR QSAR Environ Res. 2015 Jan;26(1):1-27

Part of the studies of the relation between predictive performance of (Q)SAR models for oral rat acute toxicity, and the structural features composing chemicals has been published in the peer-reviewed journal "SAR and QSAR in Environmental Research" in

January 2015 [60]. This paper reports the result of the evaluation of six (Q)SAR models for oral rat acute toxicity made within the ANTARES project.

In this part of the study of structural features to define the (Q)SAR AD, the LD50 values were considered as continuous and not classified using the ATE categorization. The dataset containing experimental LD50 values were converted in SMILES format and submitted to istChemFeat, which associated the molecules with the chemical classes defined using the functional groups. The software created a matrix reporting, for each molecule, the occurrences of each functional group and atom-centered fragment. For each of the chemical class (defined by the presence of a structural feature) the R^2 between the predictions given by the models and the experimental values were calculated. To identify the main chemical classes, they were ordered on the basis of their R^2 and for each model the ten-best (higher R^2) and ten-worst (lower R^2) were considered and compared among the five models.

17.4 Applicability domain for mutagenicity models: an a priori approach, based on chemical classes

Oral presentation by Gonella Diaza R. at 16th International Workshop on QSAR in Environmental and Health Science (QSAR2014), Milan, June 17th 2014

The possibility of defining the applicability domain of (Q)SAR models *a priori,* on the basis of experimental toxicity of "similar" molecules, was studied using a chemical classes-based approach, similar to that described for oral acute toxicity. The main idea was that chemical classes with an experimental predominant presence of a certain

effect, should be easier to predict, compared to those with a more homogeneous distribution.

As described previously, the models considered are all present within the VEGA platform, for this reason the VEGA SMILES format was chosen as the standard formalization for this part of the study. The dataset of molecules with experimental Ames test results developed within ANTARES was converted to SMILES using istMolBase. Each SMILES was associated with the mutagen / non-mutagen experimental value and the obtained dataset was submitted to istChemFeat.

The chemical classes constituted by few molecules (a threshold of ten was adopted) were not considered, whereas the other were sorted and plotted in a histogram, on the basis of their mutagens / non-mutagens distribution (evaluated by istChemFeat). In this way it was possible to identify "mutagenic" and "non-mutagenic" classes, as well as those composed by both type of chemicals. As described in the Benigni-Bossa rulebase for mutagenicity and carcinogenicity, the contemporary presence of a secondary structural alert, can inactivate the mutagenic effect of the primary one [14]. Starting from these evidences, the dataset was divided in subsets for each relevant primary chemical classes. These subsets were again classified on the basis of the presence of secondary classes, which could either enhance or quench the effect of the primary one. The same approach was than used to classify the dataset on the basis of the prediction correctness, using the outputs of the three (Q)SAR models selected as case study (CAESAR, ToxTree and SARpy).

The possibility of a chemical classes-based *a priori* definition of the applicability domain, was then studied from the comparison between the distribution of the experimental values among primary and secondary classes and the distribution of the prediction correctness.

Chapter F.
Results

18. Simple structural properties to define the applicability domain of a QSAR model for bioconcentration factor

The dataset of 860 molecules with experimentally determined values for bioconcentration factor (BCF) was predicted using the CAESAR model for BCF included in the VEGA software. No structural issues were found by VEGA in the dataset, and the results for all the molecules were saved in a text file. For each compound VEGA computed and reported the SMILES (in VEGA format), the logarithm of the bioconcentration factor (LogBCF) obtained by the two neural networks[1] composing the model and their combined value, the octanol / water partition coefficient (LogP), and the applicability domain information, showing both the global "reliability index" and all the parameters used to compute it. The same dataset was then analysed with Accelrys Discovery Studio 3.0 (DS3), which calculated the composition and molecular weight of the molecules.

The output of both software were imported and combined in Microsoft Excel to obtain a list of molecules, represented using the VEGA SMILES. Each molecule was

[1] An Artificial Neural Network (ANN) is a computational model based on the structure and functions of biological neural networks. The structure of the ANN is affected by the information that flows through it, during the initial learning stage. The results is a network composed by nodes that are associated to functions that evaluate the information provided to the network.

provided with the experimental and predicted BCF values (both expressed as LogBCF), its molecular weight and its composition, expressed as the percentage (in weight) of each atom-type. Table 4 shows an excerpt of the obtained list.

Three parameters were chosen to be assessed for the definition of the applicability domain:

- The molecular weight, since it provides a basic indication of the molecular size;
- The percentage of heteroatoms (calculated as the sum of the percentages of all atoms but carbons and hydrogens) as a representation of the "complexity";
- The percentage of halogens (calculated as the sum of Cl, Br, I and F), oxygen and nitrogen, since their electronegativity could influence the reactivity of the chemicals, and they are commonly present among organic molecules.

Table 5 reports an overview of the performance of the CAESAR BCF model on the selected dataset. As shown, the applicability domain built-in tool seems able to improve the performance of the model, even for new chemicals (not present in its training set). Regarding the molecules used to build the model, 45 of them were classified as out of the model's AD. The main reason relates to how (Q)SAR models are built. For example, while interpolating the data of the training set, the model learns the "average" behaviour of a dataset. This means that borderline molecules present in the training set, will probably not be able to give enough information, and will be outside the AD. These results will be used in the next paragraphs to compare the AD definitions obtained using the chosen properties.

Table 4. Excerpt of the table obtained from the combination of the outputs of the VEGA software for BCF and Discovery Studio 3.0. Exp. And Pred. LogBCF are the experimental and predicted (by CAESAR model) values for BCF, AD Index is the Applicability Domain Index calculated by VEGA.

ID	SMILES (VEGA)	Exp. LogBCF	Pred. LogBCF	AD Index	MW	C	H	Br	Cl	F	I	N	O	P	S	Si	Sn
1	c1cc(ccc1C(c2ccc(cc2)Cl)C(Cl)(Cl)Cl)Cl	3.60	4.19	1	354.486	0.474	0.026	0	0.5	0	0	0	0	0	0	0	0
2	O=C(O)c1c(cccc1Cl)Cl	-0.22	1.04	0.75	191.011	0.44	0.021	0	0.371	0	0	0	0.168	0	0	0	0
3	O=C(O)c1c(ccc(c1Cl)Cl)Cl	0.07	1.57	0.75	225.457	0.373	0.013	0	0.472	0	0	0	0.142	0	0	0	0
4	c1ccc2c(c1)cc3ccc4cccc5ccc2c3c45	2.73	2.98	1	252.309	0.952	0.048	0	0	0	0	0	0	0	0	0	0
5	O=[N+]([O-])c1ccc(O)c(c1)[N+](=O)[O-]	0.05	0.17	1	184.106	0.391	0.022	0	0	0	0	0.152	0.435	0	0	0	0
6	O=C(O)c1ccc(c(c1)Cl)Cl	-0.17	1.06	0.75	191.011	0.44	0.021	0	0.371	0	0	0	0.168	0	0	0	0
7	c1ccc5c(c1)ccc4cc3c(ccc2ccccc23)cc45	2.80	2.73	1	278.347	0.949	0.051	0	0	0	0	0	0	0	0	0	0
8	O(c1ccc(c(c1)C)SC)P(OC)(OC)=S	2.26	1.35	0.59	278.328	0.432	0.054	0	0	0	0	0	0.172	0.111	0.23	0	0
9	C(Cl)(Cl)(Cl)Cl	1.03	0.61	1	153.823	0.078	0	0	0.922	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Table 5. Statistics of the predictions obtained by the CAESAR BCF model on the dataset of 860 molecules. The model was evaluated on the whole dataset (global) on the molecules in common with the model's training set (in-model-training) and on the new ones (out-model-training). For each case are reported the statistics for the molecules which fall within or outside the model's applicability domain.

	Whole dataset			in-model-training			out-model-training		
	Global	In AD	Out AD	Global	In AD	Out AD	Global	In AD	Out AD
Count	860	485	375	366	321	45	494	164	330
R^2	0.63	0.79	0.46	0.82	0.84	0.63	0.47	0.69	0.39
Av. Error	0.63	0.46	0.85	0.45	0.42	0.61	0.76	0.53	0.88

18.1 Molecular weight

As explained in the previous chapter, the list of molecules was sorted and classified by their molecular weight using Microsoft Excel. A constant increment of 50 Dalton was used for the classification, with the exception of the first and last classes, which included all the molecules with MM lower than 100 Dalton and higher than 550 Dalton, respectively. Three parameters were calculated for each class:

- Number of molecules;
- Coefficient of determination (R^2) between experimental and predicted LogBCF;
- Average error, calculated as the mean of the absolute error between the experimental and predicted LogBCF values of each molecule.

Figure 2 shows the R^2 calculated for each class of molecular weight generated using the chosen thresholds. Five values of R^2 are reported for each class, which were calculated using five 10% leave-more-out iterations. These values refers to the sub-training sets. The histogram present in the upper section of the image represents the average number

of molecules (between the five sub-training sets) present in each class, in order to consider how representative each class was.

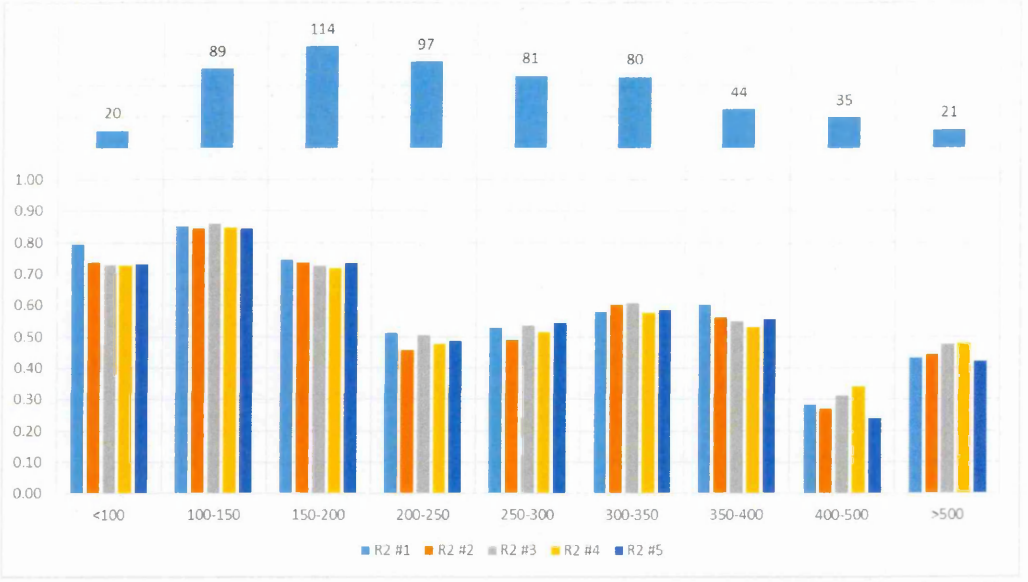


Figure 2. R2 calculated and number of molecules present in each molecular weight class. The columns in the bottom part represent the R2 calculated for the sub-training set of each leave-more-out iteration (#1 to #5). The columns in the upper part of the chart represents the average number of molecules of the five sub-training sets.

Analysing the obtained results, it seems possible to identify three "areas" in the chart (three molecular mass ranges) in which the model obtain predictions with different reliability:

- Molecules with MM smaller than 200 Dalton seems predicted with higher reliability (R^2 around 0.7-0.8);
- Molecules with MM between 200 and 400 Da seems predicted with lower reliability (R^2 around 0.5);
- Molecules with MM greater than 400 Da seems predicted with very poor reliability (R^2 around 0.2-0.4).

The identified thresholds were used to try to define the AD of the CAESAR BCF model:

- MM lower than 200 Da, the molecule is in the AD;
- MM between 200 and 400 Da, prediction could be wrong;
- MM higher than 400 Da, the molecule is out of the AD.

The rules were then tested using a prediction set and considering the model's training set (Table 6).

In all the five leave-more-out iterations the thresholds selected seems able to provide useful information for the AD definition. Molecules labelled as within the AD are consistently predicted with higher R^2 compared to "doubt" and "Out AD" ones. Using the thresholds on an external test set seems to support even more the results obtained. Considering the performance calculated on the test set, 33% of the molecules can be considered within the AD with an R^2 of 0.84. The training/test split was performed considering the whole dataset, both subsets could contain molecules which were used to build the CAESAR model, introducing a bias in the results. The dataset was therefore split using the information on the model's training set and the thresholds were applied on both subsets. As expected, the performance on new molecules (out-model-training) are lower than those on the molecules which were part of the model's training set (in-model-training). However, even in this case the MM thresholds seemed useful in the definition of the AD: 25% of the "new molecules" could be considered within the AD, and the model predicted their BCF values with a good reliability (R^2 0.7).

Table 6. Performance of the CAESAR BCF model using molecular mass thresholds to define the applicability domain (MM<=200Da, In Ad; 200Da<MM<=400Da, Doubt; MM>400Da, Out AD). The table reports the results obtained in the five iterations of the 10% leave-more-out (R^2 calculated for LMO sub-training set are also reported in Figure 2), the comparison between the training and test set, the comparison between molcules present in the model's training set and new molcules, and the performance calculated on the whole dataset.

10% leave-more-out results							
		LMO sub-training set			LMO prediction set		
		In AD	Doubt	Out AD	In AD	Doubt	Out AD
LMO-10% #1	Count	222	303	56	26	33	5
	R^2	0.79	0.55	0.33	0.81	0.62	0.48
	Average Error	0.41	0.73	1.17	0.45	0.72	1.01
LMO-10% #2	Count	219	307	55	29	29	6
	R^2	0.78	0.53	0.33	0.84	0.76	0.74
	Average Error	0.42	0.74	1.17	0.39	0.57	1.00
LMO-10% #3	Count	224	300	57	24	36	4
	R^2	0.79	0.55	0.35	0.82	0.59	0.69
	Average Error	0.41	0.73	1.13	0.50	0.75	1.45
LMO-10% #4	Count	227	299	55	21	37	6
	R^2	0.78	0.52	0.37	0.91	0.74	0.04
	Average Error	0.42	0.73	1.13	0.35	0.69	1.38
LMO-10% #5	Count	220	304	57	28	32	4
	R^2	0.78	0.55	0.30	0.83	0.57	0.09
	Average Error	0.42	0.72	1.18	0.40	0.79	0.80
Validation of the thresholds on the test set							
		Training Set			Test Set		
		In AD	Doubt	Out AD	In AD	Doubt	Out AD
	Count	248	336	61	72	121	22
	R^2	0.79	0.55	0.34	0.84	0.64	0.50
	Average Error	0.42	0.73	1.15	0.35	0.64	0.83
thresholds applied on molecules in model's training set and on new molecules							
		in-model-training			out-model-training		
		In AD	Doubt	Out AD	In AD	Doubt	Out AD
	Count	146	109	12	121	312	61
	R^2	0.85	0.77	0.66	0.70	0.44	0.33
	Average Error	0.37	0.55	0.61	0.48	0.78	1.23
thresholds applied on the whole dataset							
		In AD	Doubt	Out AD			
	Count	320	457	83			
	R^2	0.80	0.58	0.36			
	Average Error	0.40	0.71	1.07			

18.2 Percentage of heteroatoms

The software Discovery Studio 3.0 (DS3) by Accelrys, Inc. was used to determine the composition of each molecule of the dataset. DS3 calculates the percentage (in mass) of each atom-type present in a molecule, as exemplified below.

SMILES	Formula	Atom-type	n.	Atom-type weight	Total weight	Molecular Mass	% mass
c1cc(ccc1C(c2ccc(cc2)Cl)C(Cl)(Cl)Cl)Cl	$C_{14}H_9Cl_5$	C	14	12.011	168.154	354.476	47%
		H	9	1.008	9.072		3%
		Cl	5	35.45	177.25		50%

The percentages of all atoms except carbons and hydrogens were summed to obtain the percentage of heteroatoms in the molecules. This parameter was used to sort the dataset in descending order. A constant 10% decrease (from 100% to 0%) was then used to classify the molecules. The number of molecules, R^2 and average error were calculated for each class. Eleven classes were obtained, since that with the lower percentage of heteroatoms (0% to 10%) was split to consider molecules composed only by carbons and hydrogens separately from those with near 0% of heteroatoms.

Figure 3 shows the R^2 calculated on the sub-training set of the five 10% leave-more-out iterations for each class obtained. For each class, the average number of molecules between the five sub-training sets is also reported.

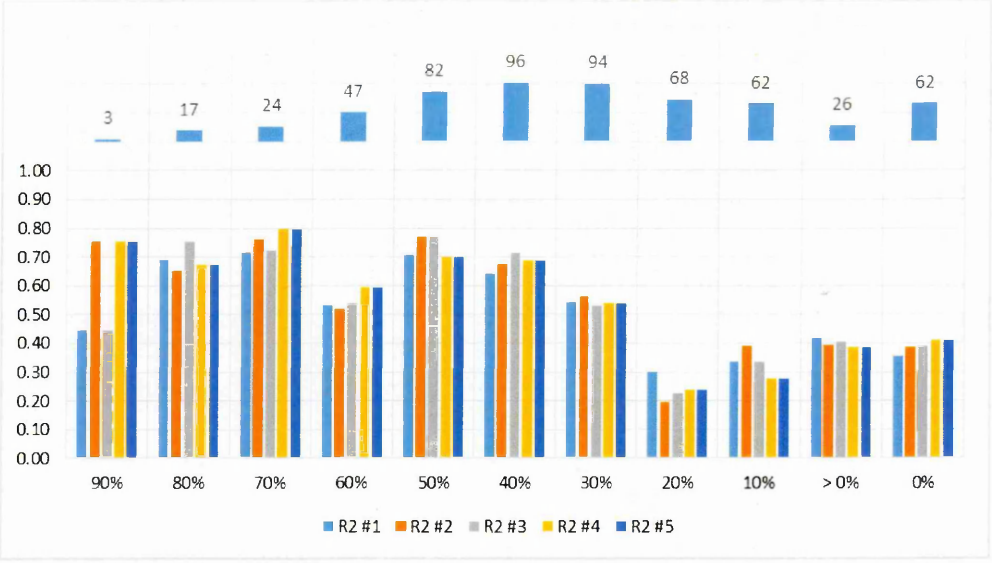


Figure 3. R2 calculated and number of molecules present in each % heteroatoms class. The columns in the bottom part represent the R2 calculated for the sub-training set of each leave-more-out iteration (#1 to #5). The columns in the upper part of the chart represents the average number of molecules of the five sub-training sets. The values on the x-axis are the lower limits of each % heteroatoms class.

Two "areas" were identified in the chart (two % of heteroatoms ranges) in which the model seemed to obtain predictions with different reliability:

- Molecules composed by more than 30% (in mass) heteroatoms seemed to be predicted with a higher reliability;
- Molecules composed by less than 30% (in mass) heteroatoms seemed to be predicted with a lower reliability.

The 30% were than used as threshold to define the applicability domain of the model. Molecules whose mass was constituted by more than 30% of heteroatoms were considered within the AD. This threshold was tested using a prediction set and considering the model's training set (Table 7).

The results do not support the use of this (probably too general) parameter as a discriminant for the definition of the applicability domain. Applying the 30% threshold to the five sub-training sets used in the leave-more-out, resulted in relatively low R^2 (0.63 to 0.67) which does not differ very much for the molecules out AD (0.50 to 0.53). Moreover, the results on the LMO prediction sets were not consistent, in one case the R^2 was even greater than that calculated on the training set. While testing this threshold on an external test set, the difference between the R^2 calculated for in AD and out AD molecules did not seem relevant (0.71 and 0.66 respectively). Furthermore, both were greater than those calculated on the training set. Finally, the R^2 did not substantially improve for molecules within AD for both molecules within the CAESAR model's training set and the "new" ones. To conclude, the 30% threshold did not seem able to provide a clear separation between reliable (in AD) and unreliable (out AD) predictions.

Table 7. Performance of the CAESAR BCF model using the 30% heteroatom threshold to define the applicability domain. The table reports the results obtained in the five iterations of the 10% leave-more-out (R^2 calculated for LMO sub-training set are also reported in Figure 3), the comparison between the training and test set, the comparison between molecules present in the model's training set and new molecules, and the performance calculated on the whole dataset.

10% leave-more-out results					
		LMO sub-training set		LMO prediction set	
		In AD	Out AD	In AD	Out AD
LMO-10% #1	Count	367	214	36	28
	R^2	0.63	0.53	0.77	0.29
	Average Error	0.68	0.60	0.64	0.66
LMO-10% #2	Count	365	216	38	26
	R^2	0.67	0.50	0.42	0.55
	Average Error	0.65	0.61	0.87	0.64
LMO-10% #3	Count	362	219	41	23
	R^2	0.67	0.51	0.42	0.40
	Average Error	0.65	0.61	0.88	0.66
LMO-10% #4	Count	359	222	44	20
	R^2	0.66	0.50	0.51	0.55
	Average Error	0.66	0.61	0.77	0.60
LMO-10% #5	Count	359	222	44	20
	R^2	0.66	0.50	0.51	0.55
	Average Error	0.66	0.61	0.77	0.60
Validation of the thresholds on the test set					
		Training Set		Test Set	
		In AD	Out AD	In AD	Out AD
	Count	403	242	144	71
	R^2	0.64	0.50	0.71	0.66
	Average Error	0.67	0.61	0.60	0.49
thresholds applied on molecules in model's training set and on new molecules					
		in-model-training		out-model-training	
		In AD	Out AD	In AD	Out AD
	Count	211	155	336	158
	R^2	0.87	0.73	0.51	0.37
	Average Error	0.43	0.47	0.80	0.70
thresholds applied on the whole dataset					
		In AD	Out AD		
	Count	547	313		
	R^2	0.66	0.53		
	Average Error	0.65	0.58		

18.3 Percentage of halogens

Using the same procedure described in the previous paragraphs, the dataset were sorted and classified by the percentage (in mass) of the halogens. In this case, molecules not containing halogen atoms were excluded since they were a large subset (475 out of 860) characterized by a R^2 very close to that calculated on the whole dataset (0.60 and 0.63 respectively). Figure 4 shows the R^2 calculated for each class obtained, on the sub-training set of the five 10% leave-more-out iterations. For each class, the average number of molecules between the five sub-training sets is also reported.

Figure 4. R2 calculated and number of molecules present in each % halogens class. The columns in the bottom part represent the R2 calculated for the sub-training set of each leave-more-out iteration (#1 to #5). The columns in the upper part of the chart represents the average number of molecules of the five sub-training sets. The values on the x-axis are the lower limits of each % halogens class, molecules without halogen atoms were excluded from the analysis.

The results suggested that, generally, the model seems to be able to provide more reliable predictions for molecules with a higher percentage of halogen atoms. Apart from the 90% class, which however is composed by only four molecules and could not

be considered reliable, molecules with percentage of halogens greater than 40% were generally better predicted. This value was chosen as a threshold for the applicability domain, however also molecules with % halogens between 30% and 20% seemed to be well predicted. The choice was made to adopt a conservative approach, excluding in this way the 30% class which seemed poorly predicted. The selected threshold was tested on an external test set and considering the molecules present in the model's training set (Table 8).

The results of five LMO runs supported the decision made. R^2s of in AD molecules of the sub-training sets were always considerably greater than those calculated for out AD ones. This was confirmed also for the prediction sets, the R^2s in this case were more variable and sometimes even greater than their respective values in the sub-training set. This was probably due to the small number of molecules considered, however the 10% sub-training/prediction split was not changed to keep this analysis consistent with those on molecular mass, heteroatoms etc. In all the cases, however, the R^2s calculated for molecules in AD were greater than those calculated for those out of AD. The selected threshold also gave good results using an external test set. The R^2 calculated for in AD molecules was probably not enthusiastic (0.65) but significantly higher than that calculated for molecules out of AD (0.25), supporting that the percentage of halogens could help in identifying poorly predicted molecules. Considering the molecules present in the model's training set separately from the "new" ones, the results were quite consistent. R^2s differed significantly between in AD and out AD molecules for both "known" and "new" molecules. The R^2 for out AD / in-model-training molecules was

0.69, which could lead to the conclusion that this method excluded a significant number of reliable predictions. As shown in Table 5 the CAESAR model performs very well on compounds within its training set (R^2 0.82), therefore even molecules predicted with a lower reliability were still well predicted. On the other hand, the performance on "new" molecules considered within the model's AD were not particularly good (R^2 0.64). However, comparing it with the results obtained on both the whole out-model-training (R^2 0.47) set and the out AD sub set (0.31), suggested that the chosen parameter could help in excluding poorly predicted molecules. Finally the results obtained on in AD and out AD subsets, defined by the AD method included in VEGA and the percentage of halogens approach, showed similar performance on both in- and out-model-training molecules.

Table 8. Performance of the CAESAR BCF model using the 40% halogens threshold to define the applicability domain. The table reports the results obtained in the five iterations of the 10% leave-more-out (R^2 calculated for LMO sub-training set are also reported in Figure 4), the comparison between the training and test set, the comparison between molecules present in the model's training set and new molecules, and the performance calculated on the whole dataset.

10% leave-more-out results					
		LMO sub-training set		LMO prediction set	
		In AD	Out AD	In AD	Out AD
LMO-10% #1	Count	121	140	15	13
	R^2	0.79	0.48	0.67	0.03
	Average Error	0.54	0.68	0.76	0.95
LMO-10% #2	Count	120	141	16	12
	R^2	0.81	0.45	0.57	0.50
	Average Error	0.54	0.70	0.80	0.76
LMO-10% #3	Count	122	139	14	14
	R^2	0.79	0.43	0.73	0.59
	Average Error	0.57	0.72	0.56	0.57
LMO-10% #4	Count	120	141	16	12
	R^2	0.77	0.44	0.90	0.38
	Average Error	0.58	0.71	0.45	0.68
LMO-10% #5	Count	121	140	15	13
	R^2	0.77	0.44	0.89	0.56
	Average Error	0.58	0.71	0.44	0.66
Validation of the thresholds on the test set					
		Training Set		Test Set	
		In AD	Out AD	In AD	Out AD
	Count	136	153	45	51
	R^2	0.78	0.44	0.65	0.25
	Average Error	0.57	0.71	0.65	0.80
thresholds applied on molecules in model's training set and on new molecules					
		in-model-training		out-model-training	
		In AD	Out AD	In AD	Out AD
	Count	96	54	85	150
	R^2	0.87	0.69	0.64	0.31
	Average Error	0.41	0.50	0.80	0.81
thresholds applied on the whole dataset					
		In AD	Out AD		
	Count	181	204		
	R^2	0.75	0.39		
	Average Error	0.59	0.73		

18.4 Percentage of oxygen and nitrogen

The last cases considered in this simple properties based approach for the determination of (Q)SAR applicability domain, considered (using the same methods described in the previous cases) the percentage of oxygen and that of nitrogen. The results obtained from the LMO showed a similar profile between oxygen and nitrogen (Figure 5). Even if there were few molecules composed by more than 40% (in mass) of both atom-types, the classes were reported to keep the representation consistent with the other case studies.

In both cases the CAESAR model seemed to obtain more reliable results in molecules devoid of either oxygen or nitrogen. The relatively high R^2 values (0.6) shown in 40% oxygen class were not considered as reliable, since related to only five molecules. For the same reason the 60% and 30% nitrogen classes were not considered. For both atom-types, the 0% threshold were used to define the model's AD, and was tested on an external test set and considering the molecules present in the model's training set (Table 9 and Table 10).

The performance evaluated in LMO supported the use of 0% oxygen as discriminant for reliable vs. unreliable prediction. Predictions for both sub-training and prediction set gave a R^2 of about 0.80 for in AD molecules and 0.40 for out AD ones, supporting the use of 0% threshold. Similar results were obtained on an external test, with R^2s even greater than those calculated for the training set, both for in and out AD molecules. On the other hand, the performance obtained using the 0% threshold for nitrogen, did not support the use of this threshold, neither in LMO nor on the external test set.

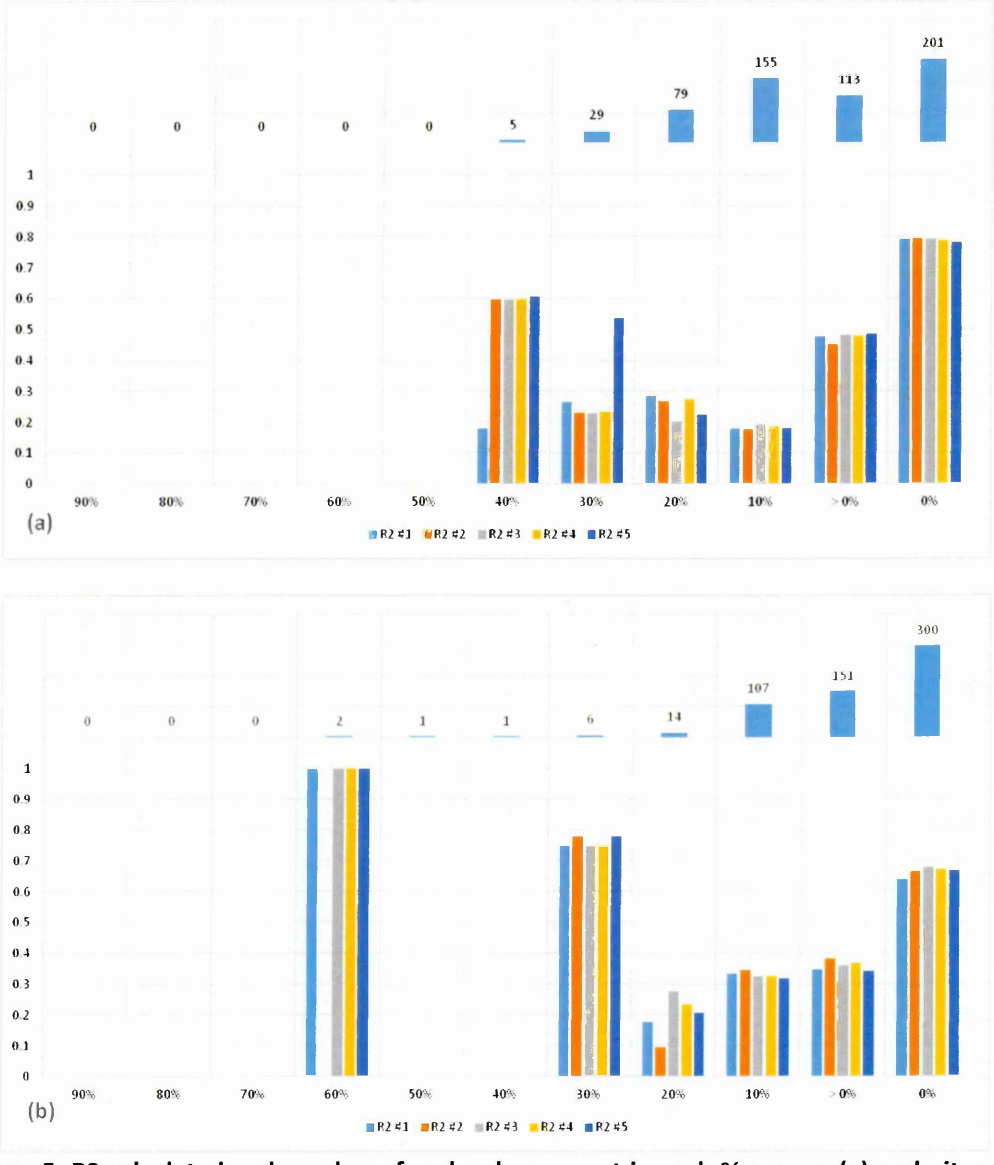


Figure 5. R2 calculated and number of molecules present in each % oxygen (a) and nitrogen (b) class. The columns in the bottom part represent the R2 calculated for the sub-training set of each leave-more-out iteration (#1 to #5). The columns in the upper part of the chart represents the average number of molecules of the five sub-training sets. The values on the x-axis are the lower limits of each % oxygen (a) and nitrogen (b) class.

Comparing the use of the oxygen threshold between molecules in and out CAESAR model's training set, the results were similar to what observed using the halogens

threshold. Even in this case, for in-mode-training subset the R^2 was higher for molecules in AD (0.90) than for those out AD (0.70). In the second case, however, the R^2 was still high. Again, this was probably due to the high performance of the CAESAR model on "known" molecules. Considering the out-model-training molecules, the use of the oxygen threshold gave interesting results, with R^2s calculated for in AD and out AD molecules differing significantly (0.69 vs. 0.37).

Even if not supported by the LMO and external test set analysis, the 0% nitrogen threshold was also applied on in- and out-model-training molecules. The results suggested that this parameter was able to discriminate between reliable and unreliable predictions obtained on known molecules, whereas did not seem to pass the most important test of recognizing good and poor predictions for "new" chemicals.

Table 9. Performance of the CAESAR BCF model using the 0% oxygen threshold to define the applicability domain. The table reports the results obtained in the five iterations of the 10% leave-more-out (R^2 calculated for LMO sub-training set are also reported in Figure 5a), the comparison between the training and test set, the comparison between molecules present in the model's training set and new molecules, and the performance calculated on the whole dataset.

10% leave-more-out results					
		LMO sub-training set		LMO prediction	
		In AD	Out AD	In AD	Out AD
LMO-10% (mean)	Count	201	380	24	40
	R^2	0.79	0.46	0.82	0.40
	Average Error	0.45	0.75	0.45	0.82

Validation of the thresholds on the test set					
		Training Set		Test Set	
		In AD	Out AD	In AD	Out AD
	Count	225	420	79	136
	R^2	0.79	0.46	0.88	0.53
	Average Error	0.45	0.76	0.37	0.68

thresholds applied on molecules in model's training set and on new molecules					
		in-model-training		out-model-training	
		In AD	Out AD	In AD	Out AD
	Count	180	186	124	370
	R^2	0.90	0.70	0.69	0.37
	Average Error	0.34	0.55	0.55	0.84

thresholds applied on the whole dataset			
	In AD	Out AD	
Count	304	556	
R^2	0.82	0.47	
Average Error	0.43	0.74	

Table 10. Performance of the CAESAR BCF model using the 0% nitrogen threshold to define the applicability domain. The table reports the results obtained in the five iterations of the 10% leave-more-out (R^2 calculated for LMO sub-training set are also reported in Figure 5b), the comparison between the training and test set, the comparison between molecules present in the model's training set and new molecules, and the performance calculated on the whole dataset.

		10% leave-more-out results			
		LMO sub-training set		LMO prediction set	
		In AD	Out AD	In AD	Out AD
LMO-10% (mean)	Count	300	281	36	28
	R^2	0.66	0.38	0.66	0.37
	Average Error	0.59	0.72	0.58	0.79

	Validation of the thresholds on the test set			
	Training Set		Test Set	
	In AD	Out AD	In AD	Out AD
Count	336	309	111	104
R^2	0.66	0.38	0.72	0.55
Average Error	0.58	0.72	0.56	0.56

	thresholds applied on molecules in model's training set and on new molecules			
	in-model-training		out-model-training	
	In AD	Out AD	In AD	Out AD
Count	248	118	199	295
R^2	0.81	0.52	0.53	0.36
Average Error	0.45	0.44	0.74	0.78

	thresholds applied on the whole dataset		
	In AD	Out AD	
Count	447	413	
R^2	0.68	0.41	
Average Error	0.58	0.68	

19. Statistical extraction of fragments related to correct or wrong predictions

The data generation for mutagenicity was performed in collaboration with the EC project ANTARES. This phase included the selection and check of the dataset to use, the selection of the (Q)SAR models to use, and the evaluation of their performance on the selected dataset. The selection of the model for mutagenicity was mainly driven by the possibility of an easy implementation of the results obtained. As already explained, the VEGA platform was developed within our research group and this would allow the exploitation of the obtained AD rules. VEGA was provided with a built-in tool for the AD check and provided predictions for both in and out AD molecules, allowing a comparison with the results obtained from this study. Moreover, the CAESAR model (implemented within VEGA) resulted as one of the best models from the ANTARES assessment [59].

The predictions obtained by the CAESAR model were used to classify the dataset for the SARpy analysis. The molecules were labelled as TP, FP, TN or FN, by comparing the predictions with the experimental Ames test values present in the dataset. The results are reported in the confusion matrix below, with the Accuracy, Sensitivity and Specificity parameters:

		Predicted	
		Mutagen	Non-Mutagen
Experimental	Mutagen	3010 (TP)	295 (FN)
	Non-Mutagen	813 (FP)	1946 (TN)

Accuracy	0.82
Sensitivity	0.91
Specificity	0.71

The two type of errors (FP and FN) were considered and modelled separately, focusing on the predicted values. The dataset of 6065 molecules was split in two subsets: the TP-FP subset, composed by 3823 molecules, and the TN-FN subset, composed of 2241 molecules. Since the SARpy tools works with the SMILES representation of molecules, they were converted to the canonical SMILES format using the OpenBabel open source software. Moreover, SARpy was designed to build SAR models; the standard procedure (as explained before) is to build a model using a training set of molecules, and testing its predictive capabilities using an external test set of new chemicals. In this case SARpy was used to model the error instead of the endpoint itself, but an external test set was needed. For this reason both TP-FP and TN-FN subsets were split in training and test sets, as explained in chapter 17. The subsets obtained are summarized in Table 11.

Table 11. Composition of the TP-FP and TN-FN subsets, and the obtained training (2/3) and test (1/3) subsets.

		Total	TP	FP	TN	FN
TP-FP	Train	2549	2007	542	0	0
	Test	1274	1003	271	0	0
TN-FN	Train	1494	0	0	1297	197
	Test	747	0	0	649	98

As explained in chapter 16, SARpy builds molecular fragments starting from the composition of the training set provided. In this phase, the user can decide the minimum and maximum number of atoms composing the fragments, the default values were used (minimum 3 and maximum 18). Once the library of fragments is built, SARpy extracts only statistically relevant ones. A simple parameter considered is the minimum number of occurrences of the fragment in the training set, the default value (minimum 3 occurrences) was used. Moreover, the statistical extraction could be performed

considering both classes of the target property (e.g. TP and FP) or focusing on only one of them. The first approach (both classes) was chosen. Again, SARpy decides if a fragment is relevant on the basis of the likelihood ratio (LR) parameter, which is usually calculated as follow:

$$LR = \frac{sensitivity}{1 - specificity} \quad [4]$$

However, SARpy fragments are only able to "predict" molecules containing them, providing only "positive" results (e.g. mutagen), including both true and false positive. If a molecule does not contain a fragment, SARpy does not predict it as negative. It is obviously possible to consider all molecules with observed negative property (e.g. non-mutagen) and that do not contain the fragment, as true negative, and observed positive molecules without the fragment, as false negative. This classification, however, would not make sense from a theoretical point of view, and would imply more calculations. The LR formulation above described, can be written in a more simple and usable way (for SARpy):

$$LR = \frac{TP}{FP} \times \frac{negative}{positive} \quad [5]$$

Where "positive" is the total number of molecules with positive observed property (e.g. molecules experimentally mutagen) and "negative" is the total number of molecules with negative observed property (e.g. experimentally non-mutagen).

Users can leave the selection of the LR threshold to SARpy, by setting if the software must maximize the coverage, minimize the error or find an optimal value. For this study,

the latter case was adopted, giving the possibility to manually analyse the impact of different LR thresholds on the model performance.

19.1 Fragments related to true positive and false positive predictions

The fragmentation of the TP molecules present in the training set, produced a total of 35422 fragments composed by three to 18 atoms. The three minimum occurrences threshold, reduced them to 6775 potential alerts. For FP molecules, SARpy produced 16845 fragments, which were reduced to 2627 potential alerts. The software compared the presence of the potential alerts between TP and FP molecules and assigned a likelihood value to each fragment. The optimisation of these values led to the final extraction of 125 fragments related to TP and 40 fragments related to FP. The complete ruleset extracted is reported in Table A - Annex A.I.

As for SAR models, these rules were used to "predict" the training set, to obtain the accuracy of prediction. In this case, however, three types of output were produced by the model: TP, FP and "none". The latter case included molecules which did not contain any of the 165 fragments. The comparison of these three classes with the two observed produced a sort of "extended" confusion matrix:

Table 12. The "extended" confusion matrix for the TP-FP training set, obtained using the TP-FP fragments extracted by SARpy.

		SARpy predictions		
		TP	FP	none
Observed	TP	1589 (T-TP)	218 (F-FP)	200
	FP	126 (F-TP)	294 (T-FP)	122

Where T-TP was the number of true positive molecules (mutagens correctly predicted by CAESAR) correctly identified as TP by SARpy, F-FP was the number of true positive molecules wrongly identified as FP by SARpy, F-TP was the number of false positive molecules (non-mutagens predicted as mutagens by CAESAR) wrongly identified as TP by SARpy, and T-FP was the number of false positive molecules correctly identified as FP by SARpy. Using these four classes an accuracy of 0.85 was calculated, suggesting that the fragments extracted by SARpy could help in identifying correct predictions. Moreover, this high accuracy was reached without leaving out too many molecules. The rules were able to cover the 87% of the training set.

To assess the performance of the ruleset on new chemicals, the test set was loaded in SARpy, which used the fragments to predict the new molecules. The results are summarized in the confusion matrix below:

Table 13. The "extended" confusion matrix for the TP-FP prediction set, obtained using the TP-FP fragments extracted by SARpy.

		SARpy predictions		
		TP	FP	none
Observed	TP	716 (T-TP)	161 (F-FP)	126
	FP	124 (F-TP)	88 (T-FP)	59

Again, using the number of T-TP, T-FP, F-TP and F-FP an accuracy of 0.74 was calculated. As expected the performance on new chemicals decreased with respect to the training

set. However, considering that this accuracy was referred to 85% of the molecules, it was considered a good result.

19.2 Fragments related to true negative and false negative predictions

SARpy produced 39248 and 7100 substructures from the fragmentation of TN and FN molecules, respectively. They were then reduced to 7384 and 911 potential alerts and finally the software extracted 78 fragments related to TN CAESAR model predictions, and 18 for FN ones. The complete ruleset extracted is reported in Table B - Annex A.I.

As previously for the TP-FP ruleset, the performance of these fragments were evaluated, generating the following extended confusion matrix:

Table 14. The "extended" confusion matrix for the TN-FN training set, obtained using the TN-FN fragments extracted by SARpy.

		SARpy predictions		
		TN	FN	none
Observed	TN	865 (T-TN)	124 (F-FN)	308
	FN	21 (F-TN)	118 (T-FN)	58

Where T-TN was the number of true negative molecules (non-mutagens correctly predicted by CAESAR) correctly identified as TN by SARpy, F-FN was the number of true negative molecules wrongly identified as FN by SARpy, F-TP was the number of false negative molecules (mutagens predicted as non-mutagens by CAESAR) wrongly identified as TN by SARpy, and T-FN was the number of false negative molecules correctly identified as FN by SARpy. From these results, an accuracy of 0.87 was calculated, covering the 76% of the training set, suggesting a possible use of these

fragments for the identification of correct and wrong predictions obtained by the CAESAR model.

The ruleset was used on to predict a prediction set of new molecules, obtaining the following results:

Table 15. The "extended" confusion matrix for the TN-FN prediction set, obtained using the TN-FN fragments extracted by SARpy.

		SARpy predictions		
		TN	FN	none
Observed	TN	356 (T-TN)	90 (F-FN)	203
	FN	43 (F-TN)	29 (T-FN)	26

The ruleset was able to predict the external set of molecules with an accuracy of 0.74, covering the 69% of the prediction set. Again, as expected the performance decreases in comparison with the training set. The coverage also decreased, in this case more significantly compared to the TP-FP case. However, thinking about using this method in combination with other AD definition, made this approach a promising one.

19.3 Analysis and application of the rules

The predictions obtained using the SARpy rulesets were analysed, counting the total number of occurrences, the number and percentage of correct predictions, and the number of wrong ones. The complete lists are reported as supplementary materials in Annex A.II. Some discrepancies between the percentage of correct assignment and the SARpy LR were observed. This was due to the software not providing all the fragments present in each molecule, but only that with the highest likelihood ratio. In fact, SARpy

uses only the "best" fragment to predict a molecule. The likelihood ratio, however, is calculated including all the molecules in which the fragment is present.

The number of occurrences of a fragment, compared to the total number of molecules in a dataset, could be of interest to decide if a fragment was relevant, and was something not considered in the calculation of the likelihood ratio. To obtain the complete list of occurrences of all the 261 structural fragments, each one was considered as a ruleset and used to predict both training and prediction set. The global results are reported as supplementary material in Annex A.II.

To use the ruleset extracted by SARpy for the definition of the applicability domain of the CAESAR model, the results shown in tables 12 - 15 were considered as follow:

- The molecule is within the AD if is predicted as TP or TN by SARpy;
- The molecules is outside of the AD it is predicted as FP or FN by SARpy;
- If a molecule does not contain any of the fragments identified by SARpy, it is not possible to determine if it can be reliably predicted or not, and another method should be used.

The accuracy, sensitivity and specificity parameters were calculated for the three classes above described. The comparison of these parameters should give an overview of the capacity of the rulesets to discriminate between reliable and unreliable CAESAR predictions. The three statistical parameters were expected to be significantly higher for molecules classified as in AD compared to those classified out AD. Table 16 reports the results of the analysis. Using the ANTARES approach, the statistical parameters were calculated considering different splits of the whole dataset:

- The SARpy training set (2/3 of the whole dataset) containing the molecules used by SARpy to obtain the ruleset;
- The SARpy prediction set, used to test the rules on "new" molecules;
- The "in model training" set, composed by the molecules in common between the whole dataset and the training set used to develop the CAESAR model;
- The "out model training" set, containing molecules "new" to the CAESAR model;
- "New" molecules for both SARpy and CAESAR model.

The latter class is the most interesting one, since it represents the "ideal" case study of a predictive model, estimating how it deals with new molecules, which were not used to build it or determine its AD.

The performance of the CAESAR model evaluated on the molecules considered as within the AD, was very high in both the whole dataset and in all of its subsets. Comparing the molecules used to build the SARpy rules (TP-FP and TN-FN training sets) with those used to validate them (TP-FP and TN-FN prediction sets), showed only a small decrease in the performance for the prediction set. The comparison of the performance between molecules in and out of the AD, suggested that the SARpy rulesets were able to discriminate between reliable and unreliable prediction. Considering the sensitivity calculated for the molecules of the SARpy prediction set, the value remains quite high (0.85) even for molecules considered as out of the applicability domain, however this was in some way expected. In fact, the analysis performed by the ANTARES project showed high performance of the CAESAR model, even for molecules not present in its training set (results reported in Table 16 – column "All molecules"). Moreover, the

ANTARES analysis suggested that the CAESAR models produces more false positives than false negatives, with a specificity value, calculated on "new molecules" of 0.60. The SARpy-based AD approach gave the best discrimination between reliable (In AD) and unreliable (Out AD) prediction exactly while considering the FP predictions. Considering again the prediction set, the specificity calculated for the molecules in the AD was significantly higher than that calculated for AD ones (0.74 vs. 0.51).

The differences between the performance obtained by CAESAR on molecules which were part of its training set and the new ones were analysed, considering the AD information. For molecules within CAESAR training set, SARpy did not seem able to identify false negative predictions. The difference in the sensitivity between in AD and out AD molecules was little (0.99 vs 0.90). Comparing the results obtained using the built-in VEGA AD tool (Table 17), suggested that the novel fragments-based AD definition was not really able to improve the VEGA tool. Regarding FP predictions, the new AD definition behaved better than on FN, with a specificity of 0.92 for in AD molecules and of 0.51 for out AD ones. Again, the VEGA AD was far more able to discriminate between reliable and unreliable predictions (0.99 vs 0.01).

Considering the molecules out of CAESAR training set, the performance obtained using the SARpy AD were slightly better than those obtained using VEGA, however the second had a greater coverage. Interestingly, the performance of the CAESAR model for molecules out of VEGA AD showed that its ability to discriminate between reliable and unreliable predictions diminish with respect to molecules within training set. The SARpy

AD showed a much more different profile between in AD and out AD molecules. However, again, the coverage in VEGA was higher.

Finally, with the aim of improving the VEGA AD built in definition, one last comparison was performed, considering molecules not used for model building nor for the definition of the applicability domain. VEGA utilizes its entire dataset to determine the AD, not just the CAESAR training set. A new subset was obtained by deleting, from the SARpy prediction set, all the molecules for which VEGA provided an experimental data. The comparison of the results obtained using the two AD approaches suggested that VEGA better discriminates between FN, whereas SARpy was more slightly more successful in eliminating FP from the in AD subset.

Table 16. Statistical analysis of the CAESAR model performance, using the SARpy rulesets for the definition of the applicability domain. For the whole dataset and each subset, the number of molecules (N.), accuracy (Acc), sensitivity (Sens) and specificity (Spec) are reported. These values have been calculated for molecules within and outside of the AD determined using the SARpy rulesets, as well as for those not containing any rules. The global values (considering In AD, Out AD and No Info together) are also reported. To provide a clear overview of the coverage of the CAESAR model, while using the AD info), the percentage of the molecules for each AD class is reported.

		All molecules	In AD	Out AD	No Info
Whole Dataset	N.	6064	3840 (63%)	1122 (19%)	1102 (18%)
	Acc	0.82	0.92	0.53	0.76
	Sens	0.91	0.97	0.72	0.80
	Spec	0.71	0.83	0.36	0.74
SARpy training set	N.	4043	2601 (64%)	754 (19%)	688 (17%)
	Acc	0.82	0.94	0.45	0.74
	Sens	0.91	0.99	0.65	0.78
	Spec	0.71	0.87	0.30	0.72
SARpy prediction set	N.	2021	1239 (61%)	368 (18%)	414 (20%)
	Acc	0.82	0.87	0.68	0.79
	Sens	0.91	0.94	0.85	0.83
	Spec	0.71	0.74	0.51	0.77
In model training set	N.	3038	1970 (65%)	478 (16%)	590 (19%)
	Acc	0.90	0.97	0.70	0.85
	Sens	0.97	0.99	0.90	0.89
	Spec	0.82	0.92	0.51	0.83
out model training set	N.	3026	1870 (62%)	644 (21%)	512 (17%)
	Acc	0.73	0.87	0.40	0.65
	Sens	0.85	0.95	0.58	0.68
	Spec	0.60	0.75	0.25	0.64
SARpy prediction set AND Not in VEGA dataset	N.	762	478 (63%)	147 (19%)	137 (18%)
	Acc	0.70	0.75	0.53	0.69
	Sens	0.84	0.89	0.74	0.72
	Spec	0.55	0.57	0.34	0.68

Table 17. Statistical analysis of the CAESAR model performance, using the applicability domain implemented within the VEGA platform.

		All molecules	In VEGA AD	Out VEGA AD
Whole Dataset	N.	6064	4902 (81%)	1162 (19%)
	Acc	0.82	0.92	0.38
	Sens	0.91	0.97	0.54
	Spec	0.71	0.86	0.28
In model training set	N.	3038	2752 (91%)	286 (9%)
	Acc	0.90	0.99	0.04
	Sens	0.97	1.00	0.13
	Spec	0.82	0.99	0.01
out model training set	N.	3026	2150 (71%)	876 (29%)
	Acc	0.73	0.83	0.49
	Sens	0.85	0.93	0.61
	Spec	0.60	0.71	0.40
SARpy prediction set AND Not in VEGA dataset	N.	762	524 (69%)	238 (31%)
	Acc	0.70	0.76	0.56
	Sens	0.84	0.91	0.67
	Spec	0.55	0.59	0.47

19.4 Fine tuning of the SARpy ruleset using different likelihood ratio thresholds

With the aim of improving the fragment-based SARpy AD's ability to discriminate between reliable and unreliable predictions, it is possible to increase the likelihood ratio thresholds, considering only more precise fragments. By doing this, the coverage is obviously destined to diminish, leaving more molecules without a decision regarding their belonging to the model's AD. The performance of different ruleset, obtained using different LR thresholds were evaluated on the training set used to extract the rules. The results are reported in Table 18.

As expected, while increasing the minimum likelihood ratio value used to consider a rule as part of the AD ruleset, the performance of the CAESAR models increased for in AD molecules and decreased for out AD ones. The separation between reliable and

unreliable predictions reached its best using rules with an infinite LR. However, in this case, for 66% of the molecules no AD info could be provided. A LR threshold of 2 helped in improving the discrimination of FP predictions between in AD and out AD molecules. The sensitivity, on the other hand did not decrease substantially for out AD molecules, with respect of the use of the whole ruleset. To obtain good results in this sense, a LR threshold of 5 could be applied. By doing this, however, a large number of molecules were unnecessarily excluded from the in AD subset.

Table 18. Performance evaluated on the SARpy training set, using different LR thresholds for selecting the relevant rules.

		All molecules	In AD	Out AD	No Info
All Rules	N.	4043	2601 (64%)	754 (19%)	688 (17%)
	Acc	0.82	0.94	0.45	0.74
	Sens	0.91	0.99	0.65	0.78
	Spec	0.71	0.87	0.30	0.72
LR >= 2	N.	4043	2107 (52%)	422 (10%)	1514 (37%)
	Acc	0.82	0.96	0.36	0.74
	Sens	0.91	0.99	0.63	0.85
	Spec	0.71	0.93	0.15	0.63
LR >= 5	N.	4043	1554 (38%)	179 (4%)	2310 (57%)
	Acc	0.82	0.99	0.21	0.75
	Sens	0.91	1.00	0.38	0.88
	Spec	0.71	0.99	0.09	0.59
LR >= 10	N.	4043	1385 (34%)	98 (2%)	2560 (63%)
	Acc	0.82	1.00	0.14	0.74
	Sens	0.91	1.00	0.17	0.89
	Spec	0.71	1.00	0.12	0.57
LR = inf	N.	4043	1326 (33%)	30 (1%)	2687 (66%)
	Acc	0.82	1.00	0.00	0.74
	Sens	0.91	1.00	0.00	0.88
	Spec	0.71	1.00	0.00	0.56

To avoid the exclusion of reliable predictions, a "two threshold" approach was applied. Since a molecule primarily associated with a TP or TN fragments were considered as in

AD, the LR threshold for these fragments was set to 2 (which gave high performance keeping a good coverage). On the other hand, to diminish the number of correct predictions in the out AD set, a threshold of 5 was set for the LR of FP and FN fragments. This reduced version of the ruleset was tested on the prediction set and the performance of the CAESAR model were also calculated considering its training set. The results are reported in Table 19.

The comparison of the performance obtained with the two rulesets, showed that using the reduced version substantially increased the discrimination between reliable and unreliable predictions. However, this improvement cost a lot in terms of coverage. For more than 40% of the molecules no information about the AD could be provided. This was not seen as a major limitation, since the idea was to integrate this approach to those already available.

Table 19. Performance of the CAESAR model evaluated using the reduced version of the SARpy ruleset.

		All molecules	In AD	Out AD	No Info
SARpy training set	N.	4043	2107 (52%)	179 (4%)	1757 (43%)
	Acc	0.82	0.96	0.21	0.70
	Sens	0.91	0.99	0.38	0.84
	Spec	0.71	0.93	0.09	0.57
SARpy prediction set	N.	2021	1020 (50%)	80 (4%)	921 (46%)
	Acc	0.82	0.88	0.51	0.78
	Sens	0.91	0.94	0.77	0.89
	Spec	0.71	0.79	0.15	0.67
In model training set	N.	3038	1589 (52%)	101 (3%)	1348 (44%)
	Acc	0.90	0.98	0.44	0.84
	Sens	0.97	0.99	0.86	0.93
	Spec	0.82	0.96	0.14	0.75
out model training set	N.	3026	1538 (51%)	158 (5%)	1330 (44%)
	Acc	0.73	0.89	0.22	0.61
	Sens	0.85	0.95	0.36	0.77
	Spec	0.60	0.81	0.09	0.47

19.5 Using the VEGA built-in AD tool to integrate the SARpy ruleset

A possibility to cover the remaining 40% of the dataset for which the SARpy ruleset is not able to provide applicability domain information, was to use those provided by VEGA itself. The performance of the CAESAR model was evaluated on the "No info" subset (2678 molecules) using the applicability domain information provided by VEGA. The results are reported in Table 20.

Comparing the results with the evaluation performed using the VEGA tools for the whole dataset (Table 17), the performance obtained was similar; the tool's ability to discriminate reliable prediction was great for molecules within the model training set, and diminished for new chemicals. Focusing on new chemicals, the ability of the VEGA tool to discriminate reliable (in AD) and unreliable (out of AD) predictions, was compared between the "no info" and the whole subset. Table 21 reports the difference calculated between the in AD and out AD values of accuracy, sensitivity and specificity.

The differences calculated for the subset of chemicals not handled by the SARpy AD ruleset were slightly higher than those calculated considering all the "new chemicals". This suggested that SARpy ruleset took care of part of the molecules for which the VEGA AD tool was less able to deal with, supporting the suggested VEGA-SARpy simple integration.

Table 20. Evaluation of the VEGA built-in AD tool on the molecules for which the SARpy reduced ruleset was not able to provide applicability domain information. The performance of the CAESAR model was evaluated on the 2678 molecules for which SARpy gave "no info" output, the applicability domain information provided by VEGA were used to determine the reliability of the predictions.

		All molecules	In AD	Out AD
Whole "No Info" dataset	N.	2678	2049 (77%)	629 (23%)
	Acc	0.73	0.87	0.25
	Sens	0.86	0.95	0.39
	Spec	0.60	0.79	0.18
In model training set	N.	1348	1146 (85%)	202 (15%)
	Acc	0.84	0.99	0.03
	Sens	0.93	0.99	0.09
	Spec	0.75	0.98	0.02
out model training set	N.	1330	903 (68%)	427 (32%)
	Acc	0.61	0.73	0.35
	Sens	0.77	0.88	0.48
	Spec	0.47	0.59	0.27
SARpy prediction set AND Not in VEGA dataset	N.	321	216 (67%)	105 (33%)
	Acc	0.62	0.69	0.46
	Sens	0.80	0.88	0.58
	Spec	0.49	0.55	0.39

Table 21. The VEGA AD tool's ability to discriminate between reliable and unreliable predictions for new chemicals. Comparison between the whole dataset and the molecules for which SARpy gave no AD information. The table reports the difference between in AD and out AD molecules (e.g. Accuracy for in AD – Accuracy for out AD).

		SARpy "no Info"	Whole dataset
out model training set	N.	1330	3026
	Acc diff	0.38	0.34
	Sens diff	0.39	0.32
	Spec diff	0.31	0.31
SARpy prediction set AND Not in VEGA dataset	N.	321	762
	Acc diff	0.24	0.20
	Sens diff	0.29	0.24
	Spec diff	0.15	0.12

The final step consisted in the "integration" of the SARpy and VEGA tools. A preliminary and simple approach was tested in this study: molecules containing TP- or TN-related fragments with a likelihood ratio of at least 2 were considered as within the applicability domain; molecules containing FP- or FN- fragments with a likelihood ratio were considered as out of the applicability domain; for molecules not associated to any fragments identified by SARpy, or containing fragments with a likelihood ratio below the selected thresholds, the applicability domain information provided by VEGA were used. The results of the performance assessment of the CAESAR mutagenicity model, using this combined AD approach, are reported in Table 22.

The results obtained were compared with the use of the VEGA AD tool alone (Table 17). Considering the whole dataset of 6064 molecules, the combined AD tool seemed able to include more correct predictions within the model AD (from 81% to 84%) without losing in accuracy (0.92 vs 0.91), sensitivity (0.97 vs 0.96), and specificity (0.86 vs 0.84). On the other hand, the smallest group of molecules excluded from the model's AD, gave lower values for all the three parameters: the accuracy decreased from 0.38 to 0.26, the sensitivity from 0.54 to 0.44, and the specificity from 0.28 to 0.16.

Molecules within the CAESAR model training set gave comparable statistical parameter for molecules included in the two ADs. The combined approach, however, seemed to exclude more good predictions from the AD: accuracy, sensitivity and specificity were higher with respect to the use of the VEGA AD alone (0.17 vs 0.04, 0.45 vs 0.13 and 0.05 vs 0.01, respectively). Finally, considering the two cases of molecules outside the model training set, and of molecules not used to build the SARpy ruleset and

not present in the VEGA dataset, the performance were comparable for those considered within the two ADs. The combined AD approach, however, seemed slightly better in excluding from the AD the wrong predictions. In this case only 18% of the molecules were excluded (VEGA excluded 31%), moreover, this smaller subset showed also decreased accuracy (0.45 vs 0.56), sensitivity (0.58 vs 0.67) and specificity (0.36 vs 0.47).

Table 22. Performance of the CAESAR model evaluated using a combination of the reduced version of the SARpy ruleset and the VEGA AD tool.

		All molecules	In AD	Out AD
Whole Dataset	N.	6064	5176 (85%)	888 (15%)
	Acc	0.82	0.91	0.26
	Sens	0.91	0.96	0.44
	Spec	0.71	0.84	0.16
In model training set	N.	3038	2735 (90%)	303 (10%)
	Acc	0.90	0.98	0.17
	Sens	0.97	0.99	0.45
	Spec	0.82	0.97	0.05
out model training set	N.	3026	2441 (81%)	585 (19%)
	Acc	0.73	0.83	0.31
	Sens	0.85	0.92	0.44
	Spec	0.60	0.72	0.23
SARpy prediction set AND Not in VEGA dataset	N.	762	624 (82%)	138 (18%)
	Acc	0.70	0.75	0.45
	Sens	0.84	0.89	0.58
	Spec	0.55	0.60	0.36

In conclusion, the results obtained suggests that modelling the errors in prediction, using a statistical method able to extract structural features related either to "good predictions" (considered within the AD) or "bad predictions" (out of the AD), could improve the definition of the (Q)SAR models applicability domain. The integration performed within this study was simple; it is likely that even better results could be obtained by integrating the structural fragments within the calculation of the VEGA

applicability domain index. Within this perspective, the likelihood ratio could play a more important role, acting for example to weight the relation "fragment present – In/Out applicability domain".

20. Using chemical classes to improve the definition of the applicability domain: a preliminary study

In: Gonella Diaza R et al. Comparison of in silico tools for evaluating rat oral acute toxicity. SAR QSAR Environ Res. 2015 Jan;26(1):1-27

A preliminary study of the possible use of chemical classes to improve the definition of (Q)SAR models' applicability domain was done for oral acute toxicity models within the ANTARES project [60]. The performance of five models were studied using a dataset of 7417 molecules, the results showed that only two of them seemed able to obtain reliable predictions (Table 23).

Table 23: Regression performance of the five models analyzed within the ANTARES project. The number and R^2 are reported for the whole dataset. T.E.S.T. provide predictions only for molecules within its AD, whereas ADMET was not able to predict one molecule. The performance obtained using the AD information (where available) are also reported.

			In model AD		Out model AD	
Model	**Predicted compounds**	**R^2**	**N.**	**R^2**	**N.**	**R^2**
ACD	7417	0.77	7299	0.78	118	0.34
T.E.S.T.	7413	0.68	7413	0.68	n/a	n/a
TerraQSAR	7417	0.64	n/a	n/a	n/a	n/a
ADMET Predictor	7416	0.54	7293	0.54	123	0.17
TOPKAT	7417	0.40	6610	0.41	807	0.28

The dataset of 7417 molecules with experimental LD50 values was analysed using istChemFeat, which identified a total of 274 chemical classes (defined by the presence

of either a functional group or an atom-centered fragment). Classes containing less than 20 molecules were removed since statistically irrelevant. The atom-centered fragments were also removed since in some cases they overlapped with the functional groups, whereas in other cases their definition was too generic. The result was a list of 105 chemical classes. For each chemical class the R^2 between the predictions given by the models and the experimental values were calculated. To identify the most relevant chemical classes, they were sorted on the basis of R^2 and for each model the ten-best (higher R^2) and ten-worst (lower R^2) were considered. Some chemical classes were present in both the ten-best and ten-worst lists of different models whereas for others the R^2 were always higher (or lower) than that calculated on the entire dataset.

20.1 Chemical classes predicted differently

Table 24 reports the four chemical classes, identified with the ten-best / ten-worst analysis, which were badly predicted by some models and well predicted by others.

The tertiary alcohols class was present among the ten-best lists of both TerraQSAR and T.E.S.T. This class also had a higher R^2 in ACD (compared to the R^2 calculated on the entire dataset), whereas the performance of ADMET Predictor did not improve. TOPKAT predicted this class with a much lower R^2 than that calculated on the entire dataset.

Sulfonates (thio-/dithio-) molecules were present in the ten-best list of TerraQSAR, whereas, as for the previous class, they were in the ten-worst class of TOPKAT. The other three software showed lower R^2 (ACD and T.E.S.T.) or, at most R^2 comparable with that calculated on the entire dataset (ADMET Predictor).

Table 24: Chemical classes present in both the ten-best and ten-worst lists of the models. The column "Identified" indicates in which lists the class was present. "Fill Gap" means that the class was not present in either ten-best or ten-worst of a model, and has been reported simply for comparison. The last column (Compare with global) indicates whether the R^2 for the class is higher (increase), lower (decrease) or nearly the same (no effect) as that calculated on the entire dataset.

Chemical Class	Identified	Model	Occurrence	R^2	Compare with global
Tertiary alcohols	Fill Gap	ACD	154	0.81	Increase
	Fill Gap	ADMET Predictor	154	0.54	No effect
	10 Best	TerraQSAR	154	0.85	Increase
	10 Best	T.E.S.T.	153	0.72	Increase
	10 Worst	TOPKAT	154	0.00	Decrease
Sulfonates (thio-/dithio-)	Fill Gap	ACD	25	0.71	Decrease
	Fill Gap	ADMET Predictor	25	0.56	No Effect
	10 Best	TerraQSAR	25	0.78	Increase
	Fill Gap	T.E.S.T.	25	0.36	Decrease
	10 Worst	TOPKAT	25	0.04	Decrease
Guanidine derivatives	Fill Gap	ACD	43	0.66	Decrease
	Fill Gap	ADMET Predictor	43	0.30	Decrease
	10 Best	TerraQSAR	43	0.76	Increase
	10 Worst	T.E.S.T.	43	0.02	Decrease
	10 Worst	TOPKAT	43	0.03	Decrease
Anhydrides (-thio)	10 Worst	ACD	34	0.37	Decrease
	10 Worst	ADMET Predictor	34	0.09	Decrease
	10 Best	TerraQSAR	34	0.76	Increase
	10 Worst	T.E.S.T.	34	0.32	Decrease
	10 Worst	TOPKAT	34	0.04	Decrease

Guanidine derivatives were again present in the ten-best list of TerraQSAR and in the ten-worst list of TOPKAT. This class was also listed among T.E.S.T.'s ten-worst and gave lower R^2 for ACD and ADMET Predictor.

Finally, the anhydrides (-thio) were present in the ten-worst list of four of the five models: ACD, ADMET Predictor, T.E.S.T. and TOPKAT. TerraQSAR seemed to give good predictions even on these molecules, which were again listed among its ten-best classes.

20.2 Chemical classes common in ten-best lists

Table 25 reports the performance of the five models, on five chemical classes which were present in the ten-best lists of at least three models but not in any ten-worst lists:

- Hydrazones showed a substantial increase of the R^2 for all the five models. This class was not listed among the ten-best classes for TerraQSAR but even in this case, the performance was considerably better compared to the global dataset.
- Sulfides showed a slight improvement of performance for the three software that list them among their ten-best classes (ADMET Predictor, T.E.S.T. and TOPKAT). ACD and TerraQSAR did not list this class among their ten-best, and the performance was comparable to the global dataset.
- Sulfoxide molecules were present in the ten-best lists of all the models, but only three show substantially better performance: ACD, TerraQSAR and T.E.S.T.
- Molecules containing trihalogenated carbons (CRX3) were predicted substantially better and were present in the ten-best lists of four models (ACD,

ADMET Predictor, T.E.S.T. and TOPKAT). Also TerraQSAR performed better compared to the whole dataset, however the difference was small.

- Imidazoles behaved similarly to previous class, but with a larger increase in the performance of TerraQSAR.

Table 25. Chemical classes present only among the ten-best lists of the models. This five classes were listed among the ten-best lists of at least three out of five models and were not present among the ten-worst list of any model. Values reported in italic were not present in the ten-best list and are reported for comparison.

	ACD		ADMET Predictor		TerraQSAR		T.E.S.T.		TOPKAT	
	no.	R^2	no.	R^2	no.	R^2	no.	R^2	no.	R^2
hydrazones	146	0.87	146	0.71	*146*	*0.71*	146	0.83	146	0.50
sulfides	*441*	*0.77*	441	0.63	*441*	*0.63*	441	0.76	441	0.50
sulfoxides	43	0.82	43	0.67	43	0.89	43	0.82	43	0.55
CRX3	349	0.83	349	0.70	*349*	*0.70*	349	0.78	349	0.54
imidazoles	252	0.85	252	0.74	*252*	*0.74*	252	0.78	252	0.55

20.3 Chemical classes common in ten-worst lists

Table 26 reports the performance of the five models on seven chemical classes present in the ten-worst lists of at least three models but not in any ten-best lists:

- Aromatic aldehydes and pyrroles were present among the ten-worst classes of ACD, ADMET Predictor, TerraQSAR and T.E.S.T. Also TOPKAT gave a very low R^2 for this class.
- Molecules containing dihalogenated carbons (CR2X2) were in the ten-worst classes of four models (ACD, ADMET Predictor, TerraQSAR and TOPKAT). Moreover, even the fifth one (T.E.S.T.) gave a very low R^2 compared to that calculated on the global dataset.

- The aromatic imines class was present in three models' ten-worst lists: ADMET Predictor, TerraQSAR and T.E.S.T. The performance of both ACD and TOPKAT models were also lower compared to the global dataset.
- Molecules containing the oxazole functional group were present in the ten-worst lists of all five models.
- Aromatic secondary amines were present in the ten-worst lists of ACD, TerraQSAR and T.E.S.T. Moreover, both ADMET Predictor and TOPKAT predicted these classes with very low R^2.
- The anhydrides (-thio) was the only class for which one model showed a R^2 value higher than that calculated for the entire dataset. The TerraQSAR model predicted this class with an R^2 of 0.76, which seemed substantially higher than for the global dataset (0.64).

Table 26. Chemical classes present only among ten-worst lists of the models. The seven classes are listed among the ten-worst lists of at least three out of five models and are not present among the ten-best list of any model. Values reported in italics and grey color, were not present in the ten-worst list and were added manually.

	ACD		ADMET Predictor		TerraQSAR		T.E.S.T.		TOPKAT	
	No.	R^2	No.	R^2	No.	R^2	No.	R^2	No.	R^2
aldehydes (aromatic)	36	0.40	36	0.00	36	0.16	36	0.04	*36*	*0.07*
anhydrides (-thio)	34	0.37	34	0.09	*34*	*0.76*	34	0.32	34	0.04
CR2X2	28	0.52	28	0.10	28	0.12	*28*	*0.34*	28	0.01
imines (aromatic)	*44*	*0.55*	44	0.00	44	0.34	44	0.06	*44*	*0.08*
oxazoles	23	0.39	23	0.02	23	0.25	23	0.01	23	0.02
pyrroles	81	0.48	81	0.04	81	0.31	81	0.26	*81*	*0.12*
secondary amines (aromatic)	174	0.48	*174*	*0.28*	174	0.21	174	0.22	*174*	*0.09*

21. A priori study of the applicability domain of (Q)SAR models using chemical classes

21.1 Identification of chemical classes related to mutagenic and "non-mutagenic" effects

The dataset of 6065 molecules with experimental Ames test values was analysed using the istChemFeat software. The software identified 133 functional groups and 107 atom-centered fragments (ACF) within the dataset. Both functional groups and ACFs were initially used as chemical classes. The complete list is reported in Table G - Annex B.III. The number of matches ranged from a single occurrence to nearly the whole dataset (for very common groups, related often to carbons or hydrogens). It was clear that groups with too few or too many occurrences were not relevant for the study. In the first case, the small number of molecules made the class statistically irrelevant, whereas in the second case the structural features were commonly present among both mutagenic and non-mutagenic molecules, as shown in Figure 6. Considering the whole dataset, the distribution of mutagens and non-mutagens was quite homogeneous, with a percentage of mutagenic molecules of 54%.

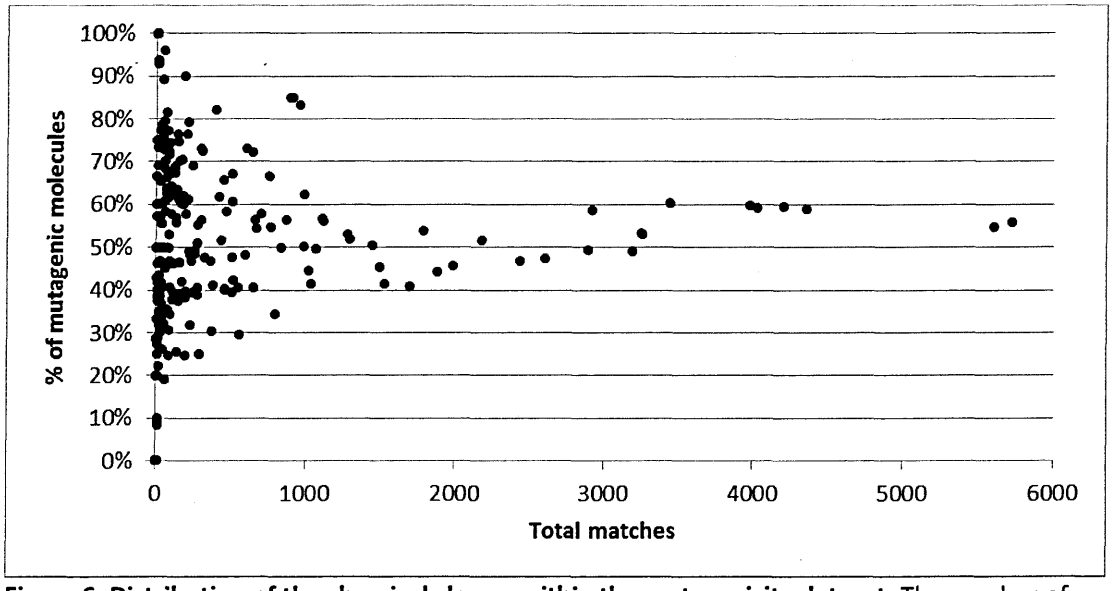


Figure 6. Distribution of the chemical classes within the mutagenicity dataset. The number of matches of each class (X axis) is reported with the % of mutagenic molecules (Y axis) present within it.

Generally, the graph showed an expected trend: very unbalanced classes were in fact not expected to cover a large part of the dataset. Classes with at least 70% of mutagenic or non-mutagenic molecules were not composed by more than 600 molecules, covering the 10% of the dataset.

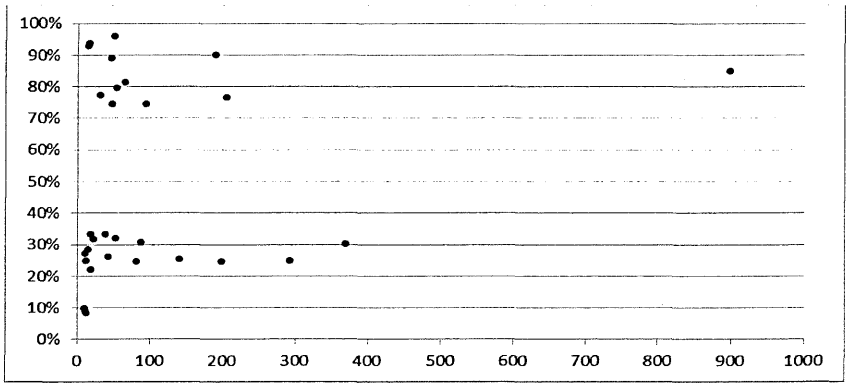
Figure 6 highlighted three "anomalies". Three groups were present in about 900 molecules, and more than 80% were mutagenic (Table 27). Actually, the classes were defined by one functional groups and two ACFs, whose definitions greatly overlapped. Moreover, the functional group was the nitroaromatic one, which is also present as a structural alert in the Benigni-Bossa ruleset for mutagenicity and carcinogenicity. For these reasons only the functional group was considered for further analysis. This decision was then extended to all ACFs, since their definitions overlapped with functional groups and/or seemed too general.

Table 27. Three chemical classes showing a percentage of mutagenic molecules higher compared to others with the same coverage.

Group	Tot Matches	Tot(muta)	% (muta)
(group no. 79) nitro groups (aromatic)	898	763	85%
(O-061) O--	958	799	83%
(N-076) Ar-NO2 / R--N(--R)--O / RO-NO	907	770	85%

Figure 7 shows chemical classes considered as possibly relevant for further analysis. As already explained ACFs were not considered; functional groups present in less than 10 or more than 1000 molecules were also excluded as well as those whose mutagenicity did not substantially differ from that of the entire dataset (a threshold of +/-20% was adopted). A total of 32 chemical classes resulted from this selection, which could possibly be used for the study of the applicability domain of (Q)SAR models for mutagenicity.

Figure 7. Chemical classes defined by functional groups present in at least 10 molecules and which showed a substantial increase or decrease of mutagenicity. The functional groups are reported below, with the total number of molecules matched, and the number and percentage of mutagenic molecules found.



Group	Tot Matches	Tot(muta)	% (muta)
(group no. 126) Aziridines	50	48	96%
(group no. 38) acyl halogenides (aliphatic)	16	15	94%
(group no. 39) acyl halogenides (aromatic)	16	15	94%
(group no. 137) Pyrazoles	14	13	93%
(group no. 74) N-nitroso groups (aliphatic)	190	171	90%
(group no. 77) nitroso groups (aromatic)	46	41	89%
(group no. 79) nitro groups (aromatic)	898	763	85%
(group no. 72) hydroxylamines (aliphatic)	65	53	82%
(group no. 140) Thiophenes	54	43	80%
(group no. 104) sulfonates (thio-/dithio-)	31	24	77%
(group no. 138) Imidazoles	205	157	77%
(group no. 73) hydroxylamines (aromatic)	94	70	74%
(group no. 116) R=CRX	47	35	74%
(group no. 17) non-terminal C(sp)	18	6	33%
(group no. 49) aldehydes (aromatic)	39	13	33%
(group no. 101) sulfonic (thio-/dithio-) acids	53	17	32%
(group no. 71) quaternary N	22	7	32%
(group no. 108) phosphates/thiophosphates	88	27	31%
(group no. 26) carboxylic acids (aliphatic)	370	112	30%
(group no. 59) oximes (aromatic)	14	4	29%
(group no. 56) imines (aliphatic)	11	3	27%
(group no. 133) Pyrrolidines	42	11	26%
(group no. 29) esters (aromatic)	141	36	26%
(group no. 4) total quaternary C(sp3)	292	73	25%
(group no. 57) imines (aromatic)	12	3	25%
(group no. 150) 1-3-5-Triazines	12	3	25%
(group no. 34) tertiary amides (aliphatic)	81	20	25%
(group no. 7) ring quaternary C(sp3)	199	49	25%
(group no. 145) Triazoles	18	4	22%
(group no. 141) Oxazoles	10	1	10%
(group no. 114) CR3X	11	1	9%
(group no. 91) anhydrides (-thio)	12	1	8%

21.2 Preliminary use of the identified classes for the applicability domain definition

To obtain preliminary information about the possible *a priori* use of chemical classes for the determination of (Q)SAR models' applicability domain, considering the distribution of the experimental data among them, a scatter plot-based analysis was performed. For each chemical classes, the percentage of experimentally mutagens and the prediction accuracy were calculated and used as X and Y values for the graph. Predictions obtained by the three models included in VEGA were considered as case studies (CAESAR, SARpy and Benigni-Bossa ruleset). The results of this analysis are reported in Figure 8. The possibly relevant chemical classes have been highlighted to get a clear overview of their predictive accuracy. Classes with less than 10 members were considered statistically not relevant and excluded from the plots.

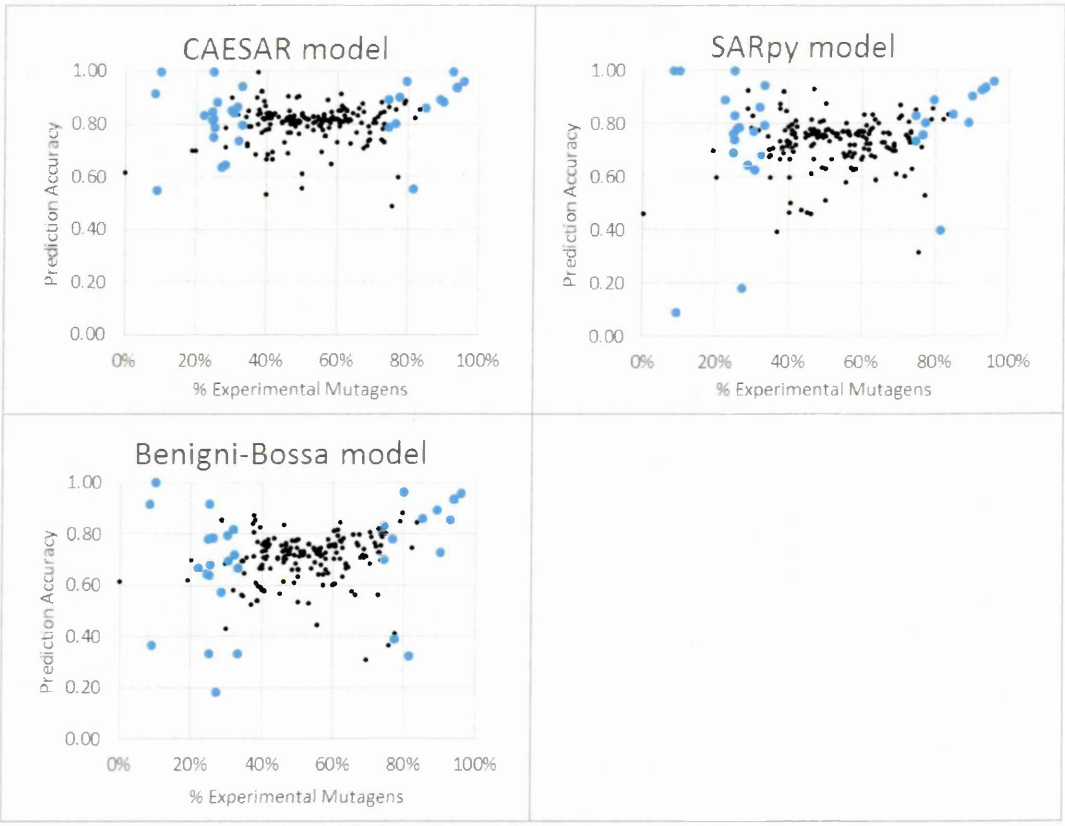


Figure 8. Preliminary analysis of the relationship between the percentage of mutagens within each chemical classes, and the accuracy of (Q)SAR models. Each point represent a chemical class and the possibly relevant classes, as previously identified, have been highlighted in blue.

To meet the hypothesis that classes with higher (or lower) ratio of mutagenic molecules should be better predicted than those with more homogeneous values, a plot with a "U" shape (typical of quadratic equations) was expected, with more "reliable classes" clustered at the two extremes of the X axis. The three plots reported in Figure 8 showed trends which seemed to support the hypothesis. However, some outliers were clearly visible, even among the classes selected as "possibly relevant for AD".

21.2.1 The CAESAR model

The plot obtained using the CAESAR's prediction accuracies showed the best distribution. This result was in line with the global performance calculated on the whole dataset (accuracy 0.82). Moreover, only four of the relevant classes were predicted with an accuracy below 0.7, and even in this case the accuracy did not decreased under 0.5 (Table 28). Three classes were composed by very few molecules (just a little above the selected threshold), the low accuracy could be therefore due to chance or to the fact that CAESAR model did not learn how to correctly predict less common molecules. The results obtained for aliphatic hydroxylamines, on the other hand, did not seem to be due to chance: 53 out of the 65 molecules in which the functional group was present, were experimentally mutagens, however CAESAR was not able to reliably predict this class. This class was further analysed and the results are reported hereinafter.

Table 28. Chemical classes identified as relevant for the definition of the applicability domain, but predicted with low accuracy by CAESAR.

Group	Tot Matches	% Mutagens	Accuracy
(group no. 72) hydroxylamines (aliphatic)	65	82%	0.55
(group no. 59) oximes (aromatic)	14	29%	0.64
(group no. 56) imines (aliphatic)	11	27%	0.64
(group no. 114) CR3X	11	9%	0.55

The possible use of the 32 chemical classes for the definition of the AD was tested and compared with the definition provided by VEGA. The performances of the CAESAR model were evaluated on the whole dataset and considering molecules within and outside the model training set, separately (Table 29). The results showed that the simple approach adopted was not really able to discriminate between reliable and unreliable predictions obtained by the CAESAR model.

Table 29. Comparison of the chemical classes-based and VEGA built-in applicability domain definitions for the CAESAR mutagenicity model.

			All molecules	In AD	Out AD
Chemical Classes AD	Whole Dataset	N.	6064	2544 (42%)	3520 (58%)
		Acc	0.82	0.84	0.80
		Sens	0.91	0.92	0.90
		Spec	0.71	0.72	0.70
	In model dataset set	N.	3038	1245 (41%)	1793 (59%)
		Acc	0.90	0.92	0.89
		Sens	0.97	0.97	0.96
		Spec	0.82	0.83	0.81
	out model dataset set	N.	3026	1299 (43%)	1727 (57%)
		Acc	0.73	0.77	0.71
		Sens	0.85	0.87	0.83
		Spec	0.60	0.62	0.59
VEGA AD	Whole Dataset	N.	6064	4902 (81%)	1162 (19%)
		Acc	0.82	0.92	0.38
		Sens	0.91	0.97	0.54
		Spec	0.71	0.86	0.28
	In model dataset set	N.	3038	2752 (91%)	286 (9%)
		Acc	0.90	0.99	0.04
		Sens	0.97	1.00	0.13
		Spec	0.82	0.99	0.01
	out model dataset set	N.	3026	2150 (71%)	876 (29%)
		Acc	0.73	0.83	0.49
		Sens	0.85	0.93	0.61
		Spec	0.60	0.71	0.40

21.2.2 The SARpy model

The performance of the SARpy model on the whole dataset was lower compared to CAESAR, with a global accuracy of 0.77, and the scatter plot obtained was in line with these lower performance. The distribution of the classes was less "compact" and a higher number of them were predicted with low accuracy. Seven of the 32 relevant chemical classes were predicted with an accuracy below 0.7 (Table 30), four of these were also badly predicted by CAESAR. Two of these "commonly" badly predicted classes

(Aliphatic amines and CR3X) were predicted by SARpy with a very low accuracy (0.18 and 0.09). Since SARpy and CAESAR were built starting from the same dataset, these results support the idea that these classes were not common within their dataset and the models did not learn how to predict them.

Table 30. Chemical classes identified as relevant for the definition of the applicability domain, but predicted with low accuracy by SARpy

Group	Occurrence TOT	% Muta	SARpy Accuracy
(group no. 72) hydroxylamines (aliphatic)	65	82%	0.40
(group no. 101) sulfonic (thio-/dithio-) acids	53	32%	0.68
(group no. 108) phosphates/thiophosphates	88	31%	0.63
(group no. 59) oximes (aromatic)	14	29%	0.64
(group no. 56) imines (aliphatic)	11	27%	0.18
(group no. 34) tertiary amides (aliphatic)	81	25%	0.69
(group no. 114) CR3X	11	9%	0.09

This hypothesis was checked using istChemFeat, resulting that only 9 aliphatic amines were present in CAESAR/SARpy dataset (7 in the training set and 2 in the test set) and the experimental mutagenicity values were also heterogeneous (5 mutagens and 4 non-mutagens). The CR2X was present only in 2 molecules (both in the training set), both of them were mutagens.

The simple AD definition using the 32 chemical classes was used to test the performance of the SARpy model, and compared to the AD definition provided by VEGA. The results obtained showed that, as for the CAESAR model, this simple approach was not able to discriminate between reliable and unreliable predictions.

Table 31. Comparison of the chemical classes-based and VEGA built-in applicability domain definitions for the SARpy mutagenicity model.

			All molecules	In AD	Out AD
Chemical Classes AD	Whole Dataset	N.	6064	2544 (42%)	3520 (58%)
		Acc	0.77	0.79	0.76
		Sens	0.82	0.85	0.79
		Spec	0.71	0.70	0.72
	in model training set	N.	3038	1245 (41%)	1793 (59%)
		Acc	0.82	0.84	0.81
		Sens	0.85	0.88	0.83
		Spec	0.79	0.78	0.79
	out model training set	N.	3026	1299 (43%)	1727 (57%)
		Acc	0.72	0.74	0.70
		Sens	0.79	0.82	0.75
		Spec	0.64	0.63	0.65
VEGA AD	Whole Dataset	N.	6064	4666 (77%)	1398 (23%)
		Acc	0.77	0.92	0.27
		Sens	0.82	0.96	0.26
		Spec	0.71	0.87	0.29
	in model training set	N.	3038	2506 (82%)	532 (18%)
		Acc	0.82	1.00	0.01
		Sens	0.85	1.00	0.02
		Spec	0.79	1.00	0.00
	out model training set	N.	3026	2160 (71%)	866 (29%)
		Acc	0.72	0.83	0.44
		Sens	0.79	0.92	0.41
		Spec	0.64	0.73	0.46

21.2.3 The Benigni-Bossa ruleset

The model based on the Benigni-Bossa (B-B) ruleset for mutagenicity gave the lowest performance on the dataset used within ANTARES, with an accuracy of 0.74. The scatter plot showed that even more of the possible relevant classes were predicted with very low accuracy. As reported in Table 32, the B-B predicted 13 of the 32 possibly relevant classes, with an accuracy below 0.70. Moreover, 5 out of the 7 relevant classes badly predicted by SARpy were also among this list (including the four badly predicted by CAESAR). As for CAESAR and SARpy, the presence of worst predicted classes were

verified within the B-B training set: only one experimentally non-mutagen aliphatic imine was present, and the same was observed for the CR3X class. 1-3-5 triazines were also badly predicted by B-B, and since only 12 molecules containing this group, their presence was also checked within the model training set, resulting in 8 molecules containing this functional group, 2 of which were mutagens and 6 non-mutagens.

Table 32. Chemical classes identified as relevant for the definition of the applicability domain, but predicted with low accuracy by the Benigni-Bossa ruleset.

Group	Occurrence TOT	% Muta	B-B Accuracy
(group no. 72) hydroxylamines (aliphatic)	65	82%	0.32
(group no. 104) sulfonates (thio-/dithio-)	31	77%	0.39
(group no. 17) non-terminal C(sp)	18	33%	0.67
(group no. 49) aldehydes (aromatic)	39	33%	0.33
(group no. 108) phosphates/thiophosphates	88	31%	0.69
(group no. 59) oximes (aromatic)	14	29%	0.57
(group no. 56) imines (aliphatic)	11	27%	0.18
(group no. 29) esters (aromatic)	141	26%	0.68
(group no. 4) total quaternary C(sp3)	292	25%	0.64
(group no. 150) 1-3-5-Triazines	12	25%	0.33
(group no. 7) ring quaternary C(sp3)	199	25%	0.64
(group no. 145) Triazoles	18	22%	0.67
(group no. 114) CR3X	11	9%	0.36

Even if the results obtained were not promising, the 32 rules were used for the definition of the applicability domain, and the B-B ruleset was evaluated on the dataset, again considering molecules within and outside its training set. The performance was compared with those obtained using the applicability domain evaluation provided by VEGA. As expected, also in this case the chemical classes-based AD simple definition was not able to provide an *a priori* definition of the applicability domain.

21.3 Considerations about the simple application of the identified classes and possible improvements

The underlying idea behind the study presented so far, was the possibility to describe an endpoint-based applicability domain, rather than a model-based one, starting from the distribution of the endpoint (mutagenicity in this case) values among chemical classes. The main advantage of this description should have been the possibility to obtain *a priori* information about chemicals which could have been difficult to predict by virtually any (Q)SAR model built for that endpoint.

Table 33. Comparison of the chemical classes-based and VEGA built-in applicability domain definitions for the Benigni-Bossa ruleset for mutagenicity.

			All molecules	In AD	Out AD
Chemical Classes AD	Whole Dataset	N.	6064	2544 (42%)	3520 (58%)
		Acc	0.74	0.76	0.73
		Sens	0.83	0.86	0.81
		Spec	0.63	0.60	0.64
	in model training set	N.	615	245 (40%)	370 (60%)
		Acc	0.77	0.76	0.78
		Sens	0.80	0.81	0.80
		Spec	0.73	0.69	0.76
	out model training set	N.	5449	2299 (42%)	3150 (58%)
		Acc	0.74	0.76	0.72
		Sens	0.84	0.87	0.81
		Spec	0.61	0.59	0.63
VEGA AD	Whole Dataset	N.	6064	3177 (52%)	2887 (48%)
		Acc	0.74	0.85	0.62
		Sens	0.83	0.91	0.74
		Spec	0.63	0.77	0.49
	in model training set	N.	615	472 (77%)	143 (23%)
		Acc	0.77	0.98	0.08
		Sens	0.80	0.97	0.16
		Spec	0.73	0.99	0.01
	out model training set	N.	5449	2705 (50%)	2744 (50%)
		Acc	0.74	0.83	0.65
		Sens	0.84	0.90	0.77
		Spec	0.61	0.73	0.52

The possible use of chemical classes was based on two main and somehow overlapping ideas: integrate simple mechanistic information in the AD definition, and consider the global (Q)SAR models as composed by several local models. In this study the chemical classes were defined by the presence of functional groups, some of them being experimentally related to mutagenic effect; for example, as reported within the Benigni-Bossa ruleset, nitroaromatic compounds generally shows a mutagenic effect. Moreover, as described in Chapter B, (Q)SAR models can be built for particular classes of molecules, rather than using large and heterogeneous datasets. The possibility to relate the AD definition to chemical classes could be a solution to integrate these two approaches, with the aim of improving the predictivity of (Q)SAR models.

A dataset of 6064 molecules was analysed, using the istChemFeat tool, to extract information about the chemical classes present and their possible relation to a mutagenic effect. As already described, the hypothesis was that if molecules sharing the same functional group had different activities, a chemical classes defined by this group could be more difficult to predict. 32 chemical classes with high percentage of mutagenic or non-mutagenic molecules were selected and highlighted in a scatter plot analysis, performed by comparing the percentage of mutagens within each classes and the accuracy of predictions of three (Q)SAR models, developed using different techniques. This analysis partially confirmed, for CAESAR and SARpy, a relation between endpoint distribution and accuracy in prediction. On the other hand, one-third of the classes selected as possibly relevant were badly predicted by the Benigni-Bossa ruleset. All the 32 classes were however used for a preliminary definition of the applicability domain of

the three models: only the molecules containing the identified functional groups were considered within the AD. The analysis of the models' performance, however, did not support this AD definition.

As described, a functional group was considered as relevant if present in at least 10 molecules and its mutagenic distribution was higher than 70%. Considering classes composed by few dozens of molecules while analysing a dataset composed by thousands of molecules could seem statistically irrelevant. However, the threshold on the distribution was thought to be able to identify less common functional groups related to mutagenic or non-mutagenic effect. Two main reasons were hypothesized for the low accuracy in predicting these less common classes: on one hand, their unbalanced distribution could be due to chance, and related only to the dataset used. On the other hand, these classes might be also uncommon within the training sets, and the model could not learn how to predict them. The three less common classes were effectively poorly represented within the models' datasets.

Another reason was thought to be related to the techniques used to build the (Q)SAR models analysed. The scatter plot analysis showed that the CAESAR model seemed more "adherent" to the hypothesis than SARpy and the Benigni-Bossa ruleset. Interestingly, CAESAR was the only model including chemical descriptors, whereas the other two were based only on structural alerts. Structural alerts-based SAR models classify molecules only on the basis of the presence or absence of certain structural fragments. It was therefore hypothesized that if a chemical class was not represented within the structural alerts, the model could be not able to correctly predict it. This could affect also CAESAR

since it was a hybrid model, composed by three models, one descriptor-based two structural alerts-based. Moreover, the structural alerts utilized were derived from the Benigni-Bossa ruleset.

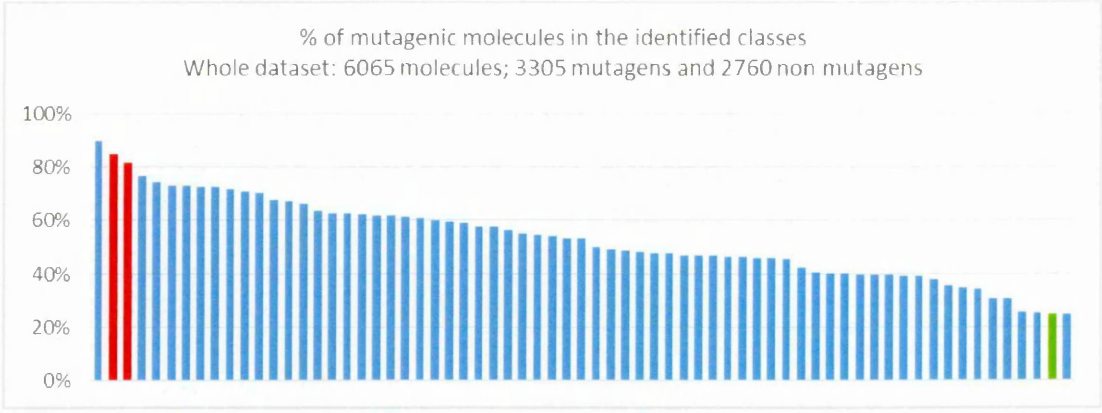
Finally, a possible improvement was considered, starting from the Benigni-Bossa ruleset. In some cases, the main rules described as related to mutagenic effect, are affected by the presence of other functional group. For example, nitroaromatic compounds are generally described as mutagenic, however the contemporary presence of a sulfonic acid or a carboxylic group decreases the mutagenic effect of the primary group. It was therefore hypothesized that considering the effect of a secondary group could improve the definition of the models' applicability domain.

21.4 Considering modulating effect of secondary functional groups

Oral presentation by Gonella Diaza R. at 16th International Workshop on QSAR in Environmental and Health Science (QSAR2014), Milan, June 17th 2014

To study the influences of the presence of a secondary functional group, three case studies were selected among the chemical classes identified by istChemFeat: nitro aromatic, aliphatic hydroxylamines and aliphatic tertiary amides. Nitro aromatic compounds were thought to be interesting due to their deviation from the global "number of matches/percentage of mutagens" trend (as shown in Figure 6). The other two classes were chosen because of their opposite possible mutagenic effect (Figure 9). As reported in Figure 7, these three classes were also identified among the 32 potentially relevant classes.

Figure 9. Mutagenicity of the three chemical classes selected as case studies for the secondary classes analysis. The histogram give a global representation of the mutagenicity of the chemical classes identified in the dataset. Nitro aromatic molecules and aliphatic hydroxylamines are highlighted in red, aliphatic tertiary amides in green.



Group	Tot Matches	Tot(muta)	% (muta)
(group no. 79) nitro groups (aromatic)	898	763	85%
(group no. 72) hydroxylamines (aliphatic)	65	53	82%
(group no. 34) tertiary amides (aliphatic)	81	20	25%

The influence of secondary classes on the mutagenic potential of the selected groups were processed with istChemFeat (complete outputs reported in Annex B.IV) and initially analysed using a scatter-plot based approach, comparing the number of molecules with the percentage of mutagens, for each secondary class (Figure 10).

The same approach used to identify the 32 primary classes was adopted also to search for secondary classes which could have a sensible influence on mutagenicity of the three classes: atom-centered fragments were excluded, a threshold of at least ten occurrences were used to exclude statistically irrelevant classes, and only secondary classes which mutagenicity differed by at least 15% compared to primary classes were considered (Figure 10 and Table 34).

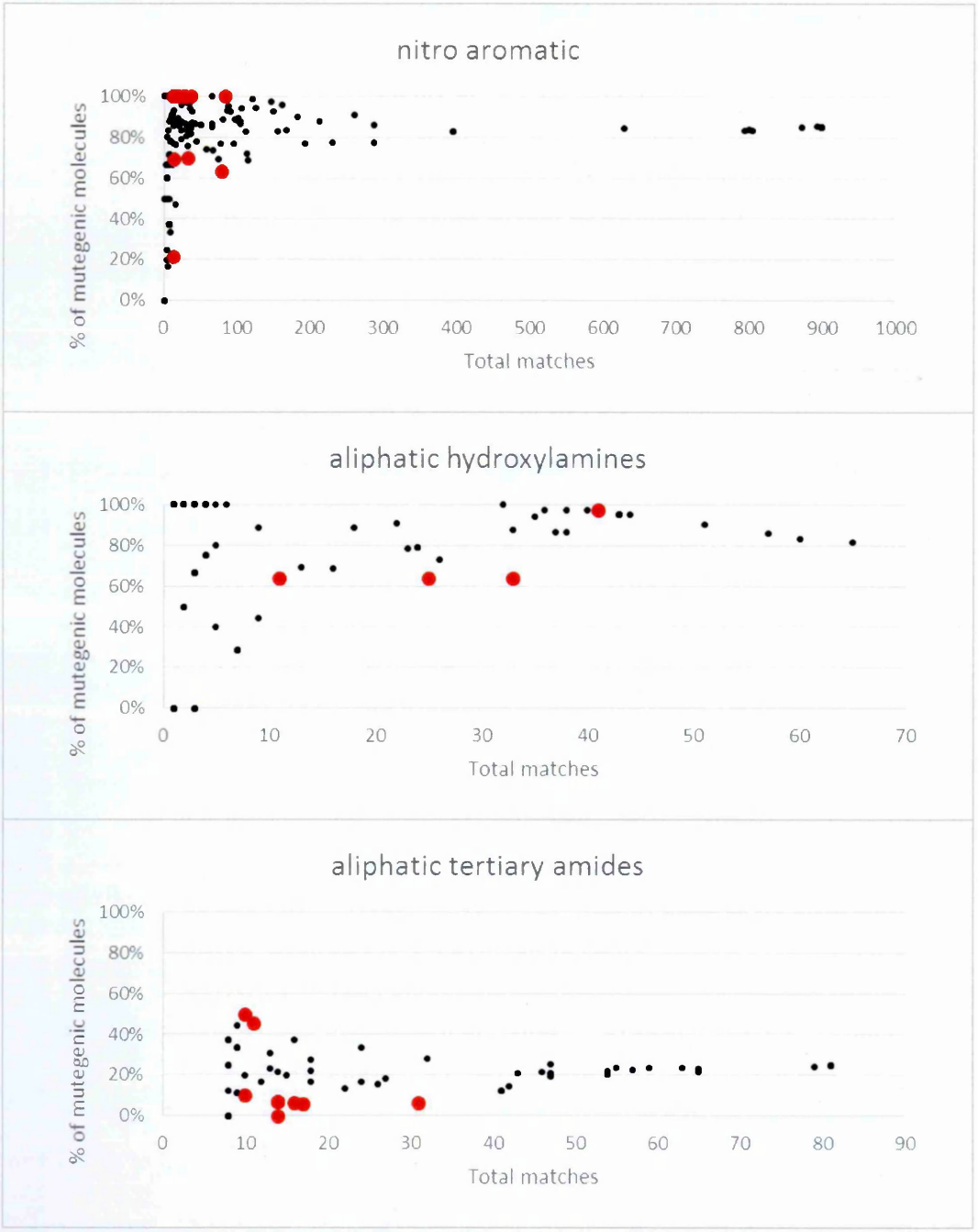


Figure 10. Distribution of the secondary chemical classes within the primary classes selected as case studies (nitro aromatic, aliphatic hydroxylamines and aliphatic tertiary amides). The number of matches of each secondary class (X axis) is reported with the % of mutagenic molecules (Y axis) present within it. Secondary classes which possibly mainly influence the mutagenicity are highlighted in red.

Table 34. Relevant secondary classes identified with istChemFeat.

Chemical Feature	Matches	Mutagens	Mutagens (%)
Nitroaromatic molecules			
(group no. 139) Furanes	84	84	100%
(group no. 140) Thiophenes	37	37	100%
(group no. 65) tertiary amines (aromatic)	29	29	100%
(group no. 33) secondary amides (aromatic)	27	27	100%
(group no. 144) Isothiazoles	19	19	100%
(group no. 64) tertiary amines (aliphatic)	14	14	100%
(group no. 148) Pyrimidines	13	13	100%
(group no. 81) hydrazones	12	12	100%
(group no. 85) secondary alcohols	33	23	70%
(group no. 26) carboxylic acids (aliphatic)	13	9	69%
(group no. 83) aromatic hydroxyls	79	50	63%
(group no. 4) total quaternary C(sp3)	14	3	21%
Aliphatic hydroxylamines			
(group no. 11) non-aromatic conjugated C(sp2)	41	40	98%
(group no. 82) hydroxyl groups	25	16	64%
(group no. 52) urea (-thio) derivatives	11	7	64%
(group no. 152) donor atoms for H-bonds (N and O)	33	21	64%
Aliphatic tertiary amides			
(C-035) R--CX..X	10	5	50%
(group no. 112) CH2RX	11	5	45%
(C-011) CR3X	10	1	10%
(group no. 95) sulfides	14	1	7%
(C-009) CHRX2	14	1	7%
(group no. 82) hydroxyl groups	31	2	6%
(group no. 26) carboxylic acids (aliphatic)	16	1	6%
(group no. 32) secondary amides (aliphatic)	17	1	6%
(O-056) alcohol	14	0	0%

The performance of the CAESAR and SARpy models, and the Benigni-Bossa ruleset were evaluated on all the secondary classes identified and compared to their mutagenicity using again a scatter plot analysis.

21.4.1 Secondary classes analysis for nitroaromatic molecules

898 nitro aromatic molecules were identified within the dataset, 85% of them were mutagens. The three models were able to obtain generally reliable predictions within

this class: the accuracy calculated for CAESAR and Benigni-Bossa was 0.86, and that calculated for SARpy was 0.83.

The scatter plot generated from the comparison between the mutagenic composition and accuracy in prediction of the secondary classes identified within the nitroaromatics subset showed a particular trend (Figure 11).

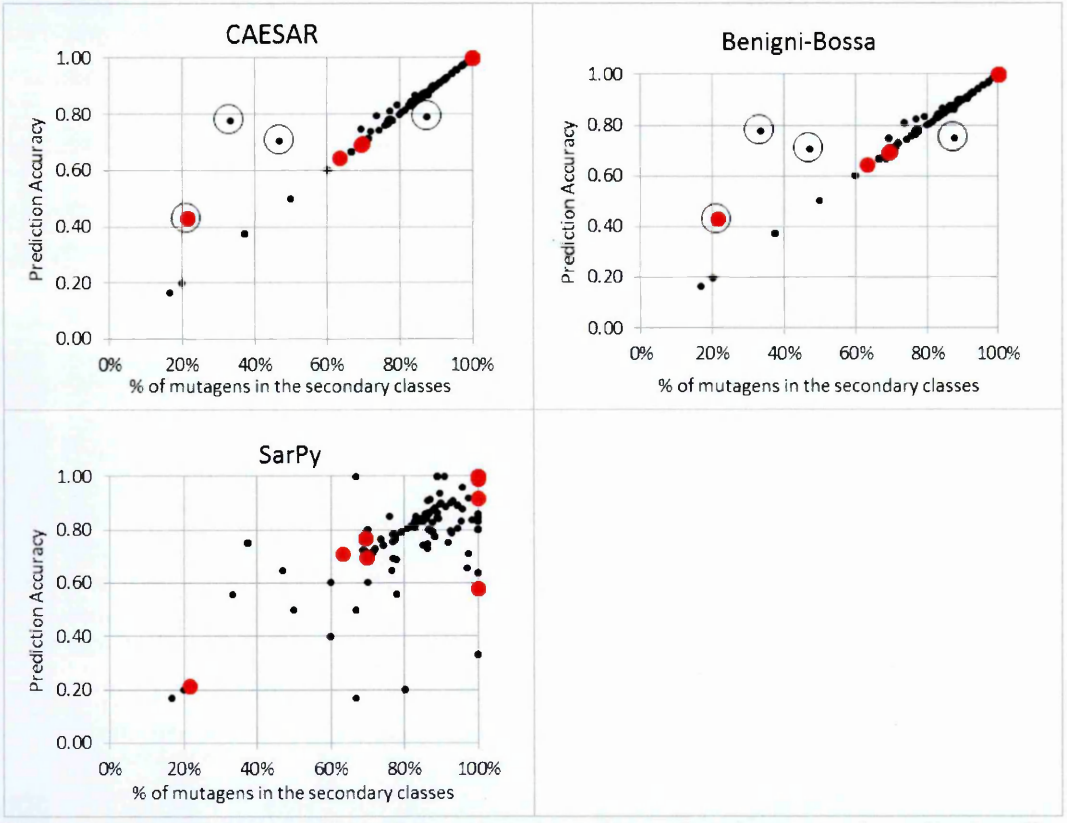


Figure 11. Scatter plot analysis of the secondary classes for nitroaromatic molecules. The classes identified as possibly relevant (Table 34) have been highlighted in red. Some classes overlaps (e.g. both Benigni-Bossa and CAESAR predicted 8 classes composed by 100% mutagens with an accuracy of 1.00). The Benigni-Bossa and CAESAR models predicted four classes with an accuracy which substantially deviated from the identified trend (circled in black). The classes with less than 5 matches are not included.

A nearly linear correlation was highlighted between the composition of the classes and the accuracy. This trend was clearer for the CAESAR model and the Benigni-Bossa ruleset, which showed a nearly identical behaviour, and was probably related to the fact the nitroaromatic functional group is described as an alert for mutagenicity within the Benigni-Bossa ruleset, and was also included in CAESAR. This means that classes with a high number of non-mutagenic nitroaromatic molecules would suffer of a high number of false positives predictions.

Four deviations from the global CAESAR/Benigni-Bossa trend were identified (Table 35). About one-third of the sulfonic acids were non-mutagens, however they were predicted by both CAESAR and Benigni-Bossa with a high accuracy (0.71 and 0.78). Similar results were obtained for a related atom-centered fragment (R-SO2-R). Carboxylic acids, on the other hand, seemed to suffer of a lower accuracy, compared to other classes with similar percentage of mutagens. These functional groups are described by Benigni-Bossa as exceptions to the nitroaromatic rule. The small number of sulfonic acids (9 molecules identified) did not allow draw significant conclusions, considering also that the two molecules badly predicted did not show the same experimental mutagenicity. The errors performed for the R-SO2-R class were also observed. Nitroaromatic compounds bearing this ACF did not show an experimental clear behavior, however they were mainly predicted as mutagens (4 errors out of 5 were false positives). Similar results were obtained from the analysis of the carboxylic acids class. In this case, 6 molecules were badly predicted by the Benigni-Bossa ruleset (5 of them also badly predicted by CAESAR). Also in this case the errors were equally

distributed among experimentally mutagens and non-mutagens (3 were false positive and 3 false negative). These results suggested that the applicability of the main nitroaromatic rule in combination with the either sulfonic acid or the carboxylic exceptions would need more investigations (not possible within this study due to lack of data).

Table 35. Secondary classes present within the nitroaromatic subset, which deviated substantially from the global Benigni-Bossa (B-B)/CAESAR trends

group	Occurrence	% Mutagens	B-B Accuracy	CAESAR Accuracy
(group no. 27) carboxylic acids (aromatic)	24	88%	0.75	0.79
(group no. 101) sulfonic (thio-/dithio-) acids	9	33%	0.78	0.71
(S-110) R-SO2-R	17	47%	0.71	0.78
(group no. 4) total quaternary C(sp3)	14	21%	0.43	0.43

The SARpy model is also based on structural alerts, however these rules were derived statistically from its training set. The results suggested that the nitroaromatic functional group was present among the model's rules, however probably as part of one or more "bigger" fragments. For this reason the correlation between the classes' mutagenic propensity and the accuracy in prediction presented more "outliers". The isothiazoles was the only "deviated" class which was previously listed among the possibly relevant ones. The 19 molecules composed by both a nitroaromatic and an isothiazolic group were mutagens, however half of them were predicted as non-mutagens by SARpy. Also in this case the number of molecules seemed not sufficient to establish a clear applicability domain rule, however the contemporary presence of these two groups could be a potential target for future investigations.

21.4.2 Secondary classes analysis for aliphatic hydroxylamines

The aliphatic hydroxylamines functional group matched for 65 molecules within the dataset used for this study, the 82% of them were experimentally mutagens. None of the analysed models seemed able to obtain reliable predictions: the accuracy calculated were 0.54 (CAESAR), 0.40 (SARpy) and 0.32 (Benigni-Bossa).

The scatter plot analysis performed by comparing the mutagenicity composition of the secondary classes with the prediction accuracy of the three models, showed a clear trend (Figure 12). Secondary classes characterized by a high percentage of mutagens were generally predicted with a low accuracy, whereas better results were obtained for classes with the opposite composition. This trend was clearer for the Benigni-Bossa ruleset than for both SARpy and CAESAR, which showed some deviations (Table 36).

The Benigni-Bossa ruleset did not list aliphatic hydroxylamines among its alerts for mutagenicity, this could explain the identified trend. Molecules belonging to this class were generally predicted as non-mutagens, therefore secondary classes richer in mutagenic molecules were badly predicted, producing a high number of false negatives.

The CAESAR model, as explained, is a hybrid model. Apart from including some of the Benigni-Bossa structural alerts, CAESAR includes a molecular descriptor-based model. The scatter plot obtained from the CAESAR prediction showed that the secondary classes are generally better predicted compared to Benigni-Bossa, suggesting that the CESAR model could have partially learned how to predict aliphatic hydroxylamines, through its molecular descriptors.

The analysis of the results obtained with SARpy seemed to support the idea that even this model was not able to correctly predict mutagenic hydroxylamines. Interestingly, the secondary classes with a percentage of mutagenic molecules between 60% and 80% were predicted with an accuracy similar to that obtained with CAESAR. On the other hand, the accuracy for secondary classes composed by more than 80% of mutagens were predicted with an accuracy closer to that obtained by Benigni-Bossa. Since SARpy and the descriptor-based model included in CAESAR have been developed using the same dataset, SARpy could have also learn, this time through its structural alert, how to partially better predict aliphatic hydroxylamines.

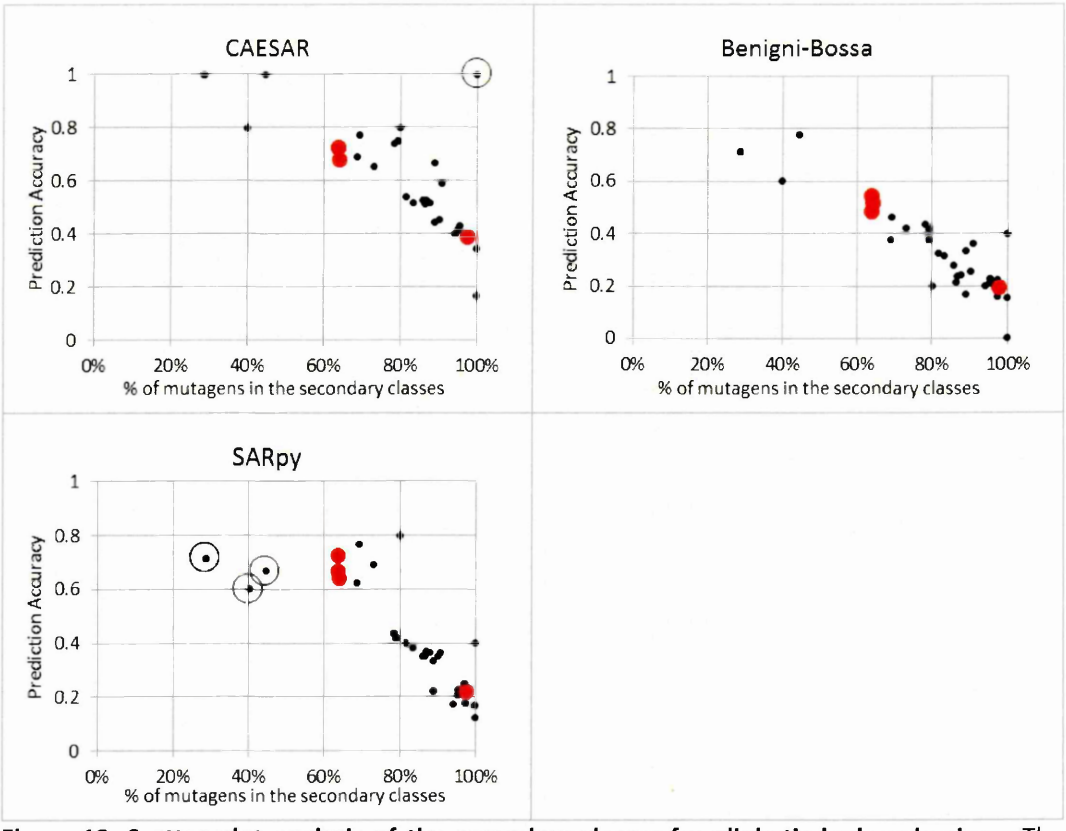


Figure 12. Scatter plot analysis of the secondary classes for aliphatic hydroxylamines. The classes identified as possibly relevant (Table 34) have been highlighted in red, whereas those deviating from the global trend are circled in black. The classes with less than 5 matches are not included.

Table 36. Secondary classes present within the aliphatic hydroxylamines subset, which deviated substantially from the global prediction trends.

Group	Matches	EXP	CAESAR Accuracy
	CAESAR		
(C-008) CHR2X	5	100%	1.00
	SARpy		
(group no. 5) ring secondary C(sp3)	9	44%	0.67
(O-057) phenol / enol / carboxyl OH	5	40%	0.60
(group no. 3) total tertiary C(sp3)	7	29%	0.71
(group no. 6) ring tertiary C(sp3)	7	29%	0.71
(C-011) CR3X	7	29%	0.71

21.4.3 Secondary classes analysis for aliphatic tertiary amides

The aliphatic tertiary amides group matched for 81 molecules within the dataset used for this study, 25% of them were experimentally mutagens. CAESAR and Benigni-Bossa predicted this class with very good accuracy (0.81 and 0.78), whereas SARpy performed worse (0.69).

The scatter plot analysis highlighted a trend between the percentage of mutagens within the secondary classes and the accuracy in prediction of the Benigni-Bossa ruleset (Figure 13). The model predicted secondary classes with a low presence of mutagenic molecules with a higher reliability compared to those with increasing number of mutagens. Also the CAESAR model seemed to be characterized by a similar trend, highlighted by the position of the more relevant classes (previously reported in Table 34). The SARpy model, which displayed the worst accuracy for the aliphatic tertiary amides class, did not seem to follow any particular trend, predicting the secondary classes with an accuracy independent from their mutagenic composition.

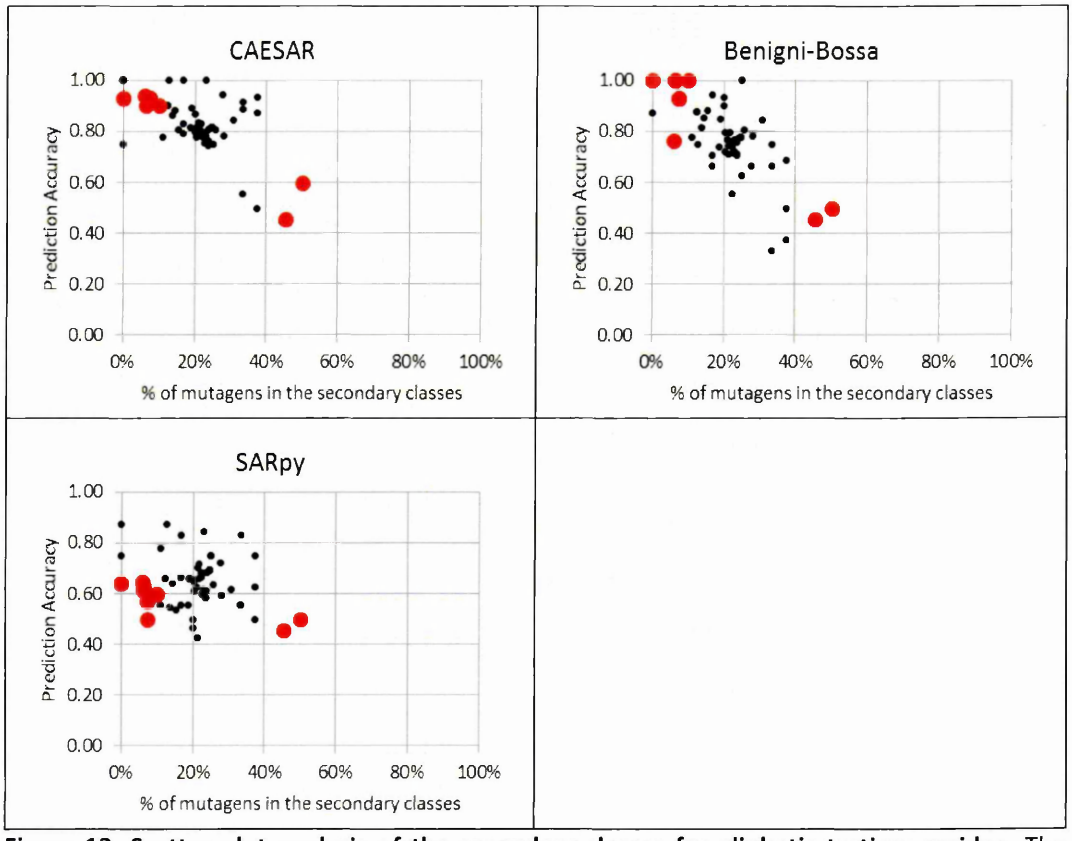


Figure 13. Scatter plot analysis of the secondary classes for aliphatic tertiary amides. The classes identified as possibly relevant (Table 34) have been highlighted in red.

Considering again the models' algorithms, neither Benigni-Bossa nor CAESAR were provided with functional groups potentially related to non-mutagenic effect of aliphatic tertiary amides. On the other hand, Benigni-Bossa include a structural alert for *α, β unsaturated carbonyls* (SA10) which could in some cases overlap with the aliphatic tertiary amides definition (Figure 14). SA10 was described by Benigni-Bossa as an alert with common carcinogenicity effect. In their study, 29 out of 38 molecules matched were experimental carcinogens. This relation was less strong when applied to mutagenicity, only 8 out of the 26 molecules that matched this structural alert were mutagens. The number of aliphatic tertiary amides that fired SA10 within the dataset

used in this study was 10, all of them therefore predicted as mutagens, however only half of them were experimentally mutagens.

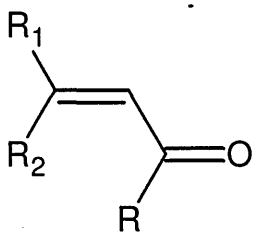
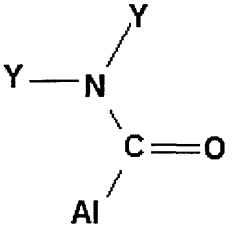
α, β unsaturated carbonyls (SA10)	aliphatic tertiary amides
	

Figure 14. Representation of the Benigni-Bossa SA10 and of the aliphatic tertiary amides class. R1 and R2 can be any atom/group, except alkyl chains with more than carbons or aromatic rings. R can be any atom/group, except OH and O-. Y can be any aliphatic or aromatic atom, except hydrogen or C=O. Al can be hydrogen or aliphatic group linked through carbon.

The results obtained suggested that for chemical classes defined by the presence of a functional group not identified as a structural alert by SAR models, the mutagenicity of secondary classes could provide information about the prediction accuracy of structural alerts-based models. This results also supported the initial hypothesis of a possible relation between the mutagenic composition of chemical classes and the performance of models. Coming back to the global scatter plot obtained for the primary classes (Figure 8), a trend was expected and partially confirmed: classes with very unbalanced distribution of the mutagenic effect could be better predicted by models. The scatter plot obtained for aliphatic tertiary amides seemed to represent "half" of the expected trend. This was due to this class mainly composed by non-mutagenic molecules, therefore no secondary classes with a very high presence of mutagens were identified.

22. Development of a novel tool for the automatic extraction of primary and secondary classes

The approach afore described has been further studied within our research group, with the aim of identifying rules for mutagenicity. The results obtained from the study presented within this thesis and those obtained from the identification of mutagenicity-related rules, led to the development of a novel tool able to automatically identify relevant primary and secondary classes. This novel tool, called istRex, is still under development and beta testing. The list of chemical classes integrated within istRex has been improved compared to that included within istChemFeat. The number of atom-centered fragments have been reduced, keeping only those most relevant and that do not overlap with functional groups. The manual analysis performed within our group showed that two functional groups present on the same aromatic ring seemed to influence the mutagenicity of the molecule depending on their relative position (ortho, meta or para). For this reason a new set of classes were included, which described the relative position of atoms such as oxygen, nitrogen, etc. A further improvement compared to istChemFeat consists in the adoption of the p-value to extract only relevant classes. IstRex requires the user to choose a target value (e.g. mutagen or non-mutagen) and extract only classes with a composition that substantially differs from that of the entire dataset. The p-value is used to determine if this difference can be related to chance or is relevant. Once the primary classes are identified, istRex performs the same analysis on each subset, to identify secondary rules. Unlike istChemFeat, the new tool consider also the occurrence of the functional group within the molecule.

The novel tool developed was used for a preliminary automated analysis of the mutagenicity dataset. Primary and secondary classes relevant to experimental mutagenicity and accuracy in prediction of CAESAR, Benigni-Bossa ruleset and SARpy were extracted and compared. To identify classes predicted with a substantially different accuracy, the molecules were submitted to istRex, using correct/wrong as property, and targeting the correct predictions. The results were analyzed using the same scatter plot approach afore described (Figure 15).

The scatter plot analysis of the CAESAR accuracy, performed considering only the primary classes, showed that most of those identified as relevant for their particularly high accuracy were in common with those identified to increase or decrease mutagenicity (green dots in Figure 15). Only two classes were characterized by both a low accuracy and an unbalanced mutagenicity composition (red dots): aliphatic hydroxylamines (65 molecules, 82% mutagens, accuracy 0.54) and CR3X (11 molecules, 9% mutagens, accuracy 0.36). These two functional groups were also identified in the previous manual analysis. The accuracy analysis also identified 4 classes which were not identified on the basis of the mutagenic composition (green and red triangles). Two of these classes showed a particularly low accuracy: aliphatic imines (11 molecules, 27% mutagens, accuracy 0.27) and molecules containing selenium (3 molecules, 100% mutagens, accuracy 0.00).

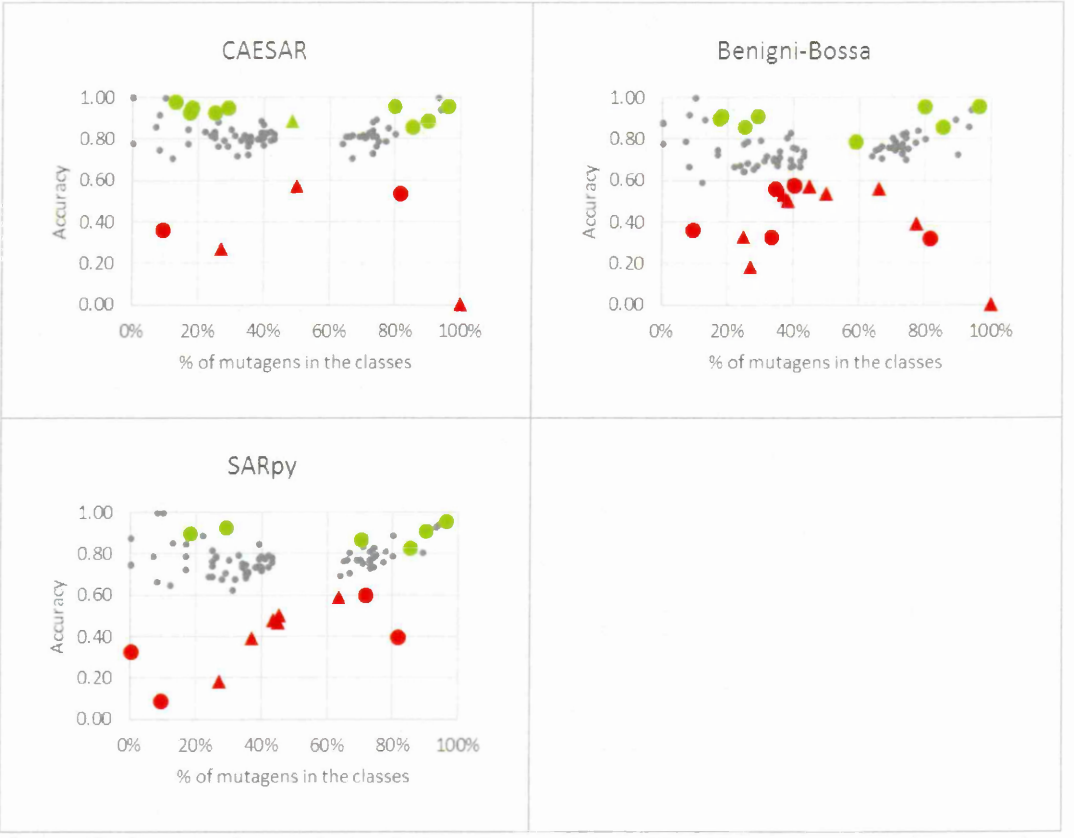


Figure 15. Scatter plot analysis of the primary classes identified by istRex. The grey dots represent classes which were identified as relevant on the basis of their experimental mutagenicity, but were not identified on the basis of prediction correctness. Bigger dots represent chemical classes identified both as relevant for mutagenicity and prediction correctness, whereas the triangles represent those identified only for accuracy. The color green means that the accuracy for the classes was greater compared to the global dataset, whereas the meaning of the red color is the opposite.

Both the Benigni-Bossa ruleset and the SARpy model predicted, compared to CAESAR, a higher number of classes characterized by high or low presence of mutagenic molecules. These results confirmed what already obtained from the manual analysis afore described, and led to idea that the chemical classes-based approach for the determination of the applicability domain seems to perform better for models which

include molecular descriptors. The applicability domain of models including only structural alerts seemed less suitable to be defined using this approach due to the possible lack of information about the chemical classes within the structural alerts.

Chemical classes with unbalanced mutagenic composition which were predicted with an accuracy substantially lower compared to that calculated on the whole dataset were further analyzed (Table 37). The aliphatic hydroxylaminic group, as afore described, seemed to be related to mutagenic effect. Neither of the three models used were however able to predict this class with an acceptable accuracy. The analysis of the secondary classes showed, for the CAESAR model, a possible relation between the presence of aromatic rings and the low accuracy. Four secondary classes were identified to substantially decrease (NON-TARGET direction) the accuracy of aliphatic hydroxylamines, all of them related to the presence of aromatic rings. Considering aliphatic hydroxylamines that did not matched the secondary rules, 41 molecules remained, predicted by CAESAR with an accuracy of 0.71. IstRex extracted a secondary rule for aliphatic hydroxylamines also for the SARpy model. In this case, molecules composed by less than 1 nitrogen atom were predicted with an accuracy of 0.15. Considering hydroxylamines composed by more than 1 nitrogen atom resulted in 26 molecules, predicted with an accuracy of 0.77.

Table 37. Primary classes identified with high or low percentage of mutagens but predicted with low accuracy. These primary classes were identfied (for each model) both as substantially more or less mutagens compared to the whole dataset, but predicted with a low accuracy (compared to that for the whole dataset). The secondary classes identified either from the mutagenic or the accuracy analysis are also reported.

				Model Prediction		Experimental values	
	Primary classes	**Secondary classes**	**Matches**	**Rule direction**	**Accuracy**	**Rule direction**	**% Mutagens**
CAESAR	hydroxylamines (aliphatic) >= 1		65	NON-TARGET	0.54	TARGET	0.82
		aromatic C(sp2) >= 8	24	NON-TARGET	0.21		
		aromatic (from 3 to 6 membered) rings >= 2	24	NON-TARGET	0.21		
		unsubstituted benzene C(sp2) >= 6	22	NON-TARGET	0.14		
		benzene rings >= 2	22	NON-TARGET	0.14		
	CR3X >= 1		11	NON-TARGET	0.36	NON-TARGET	0.09
Benigni-Bossa	aliphatic tertiary C(sp2) >= 1		458	NON-TARGET	0.58	NON-TARGET	0.4
		CHRX2 >= 1	15	TARGET	1	TARGET	1
		ketones (aliphatic) >= 1	102	NON-TARGET	0.41		
		C-Halogen in ortho position, on any ring >= 1	6	NON-TARGET	0		
		O-Halogen in para position, on any ring >= 1				TARGET	0.89
		nCl >= 1				TARGET	0.7

				Model Prediction		Experimental values	
	Primary classes	**Secondary classes**	**Matches**	**Rule direction**	**Accuracy**	**Rule direction**	**% Mutagens**
	CRX3 >= 1		84	NON-TARGET	0.56	NON-TARGET	0.35
	CR3X >= 1		11	NON-TARGET	0.36	NON-TARGET	0.09
	aldehydes (aromatic) >= 1		39	NON-TARGET	0.33	NON-TARGET	0.33
	hydroxylamines (aliphatic) >= 1		65	NON-TARGET	0.32	TARGET	0.82
SARpy	Thiazoles >= 1		81	NON-TARGET	0.6	TARGET	0.72
	hydroxylamines (aliphatic) >= 1		65	NON-TARGET	0.4	TARGET	0.82
		nN < 2	39	NON-TARGET	0.15		
	CR2X2 >= 1		9	NON-TARGET	0.33	NON-TARGET	0
	CR3X >= 1		11	NON-TARGET	0.09	NON-TARGET	0.09

Part III – Discussion, conclusion and future perspective

Chapter G.

Structural properties, functional groups and molecular fragments: are they able to define (Q)SAR models' applicability domain?

The applicability domain of the CAESAR model for bioconcentration factor was chosen as a case study to analyse the possible use of simple properties to determine its applicability domain. Bioconcentration factor is a relatively simple endpoint to model, which usually depends for example on molecular size and the octanol/water partition coefficient. Such an endpoint was thought to benefit from the use of simple properties to determine the AD. Three main aspects were considered for the determination of the AD of the CAESAR model: the molecular size, complexity and the electronegativity. These aspects were represented, within this study, using three simple properties: molecular weight, atomic composition, and presence of polar atoms (halogens, oxygen and nitrogen were considered).

Molecular weight and halogens' presence gave the best results. The R^2s calculated for molecules included in the model AD substantially differed from those calculated on the excluded ones, suggesting that both the parameters adopted provided important information about the reliability of the model's predictions. Two molecular weight thresholds were adopted to define the AD: 200 Da and 400 Da. Molecules with a MW

lower than 200 were included in the AD, whereas those with a MW higher than 400 were excluded. No applicability domain information could be obtained for molecules with MW between these thresholds. The comparison of this definition with the applicability domain information provided with the CAESAR predictions showed that the molecular weight approach discriminated slightly better between reliable and unreliable predictions. Considering only the molecules not present within the training set used to build the CAESAR model, the R^2 calculated for molecules within the MW-based AD was comparable to that calculated using the CAESAR AD (0.70 and 0.69). Unreliable predictions were however better identified by the MW approach (R^2 0.33), the R^2 for molecules out of the CAESAR AD was in fact higher (0.39).

The presence of halogens seemed also to affect the CAESAR predictions' reliability. Molecules whose mass was composed by more than 40% of halogen atoms were generally better predicted by CAESAR. This value was therefore used to define the model's AD. Indeed, a second threshold was also applied since a high number of non-halogenated molecules were present and the R^2 calculated in preliminary analysis did not differ from that calculated on the whole dataset. Therefore, non-halogens were not considered and another AD approach should be used for their characterization. Considering the case of "new chemicals" (not included in CAESAR training set), the adopted threshold provided good information for the AD definition: the R^2s calculated for in AD and out AD molecules were 0.64 and 0.31. These results were also comparable with those obtained using the AD information provided by the model.

Among the other parameters analysed, only the presence of oxygen seemed to affect the CAESAR performance. In the preliminary analysis it was observed that molecules not containing oxygen were generally better predicted by CAESAR, therefore a threshold of 0% was tested. The performance evaluated for in AD and out AD molecules, considering only molecules not present in the CAESAR training set, gave results comparable to those obtained with the halogens-based approach. Molecules which did not include any oxygen (in AD) were predicted with a R^2 of 0.69 whereas a R^2 of 0.37 was obtained for molecules outside the applicability domain.

The results suggested that the CAESAR model could benefit from the selected thresholds. It is important to underline that the aim was not to find the best method but to study approaches able to improve the current AD definition. The CAESAR model is implemented within the open source VEGA platform developed by our research group, and the applicability domain information are currently provided by a built-in tool. This will simplify the study of the possible integration of the identified threshold.

A preliminary study of the described approach was performed for a more complex endpoint: rat oral acute toxicity. However, the results obtained (not reported within this thesis) did not support the use of simple properties for the characterization of the applicability domain of the models studied. It was hypothesized that such simple properties were too general to be applied on such a complex endpoint. Therefore, rather than considering the presence of single atoms, a chemical classes-based approach was studied.

Using the freely available istChemFeat software, the molecules were classified on the basis of their functional groups, obtaining 105 chemical classes. The evaluation of the predicting performance of (Q)SAR models for the identified classes was performed in collaboration with the ANTARES project. The performance of five models for oral rat acute toxicity were being analysed within this project, using a dataset of more than 7000 molecules with experimental LD50 values. These models were implemented within four commercial software: ACD/Labs ToxSuite, Simulation Plus ADMET Predictor, TerraBase Inc. TerraQSAR and Accelrys TOPKAT. Only one model was freely available, as implemented within the U.S. EPA Toxicity Estimation Software Tool (T.E.S.T.).

The main aim of the chemical classes-based study was to obtain applicability domain information related to the endpoint rather than on each single models. Ten classes predicted with the highest R^2 were considered for each model. The comparison of these classes resulted in the identification of five classes predicted with a R^2 higher than that calculated on the whole dataset (Table 25 - Chapter F), suggesting their possible application as rules for an endpoint-based applicability domain. Moreover, the analysis performed within the ANTARES project suggested that the applicability domain information provided by the models were probably able to identify only a small portion of the unreliable predictions.

The ACD software, which showed the best overall performance, identified 118 molecules as out of its AD. The R^2 calculated on these molecules was 0.34, whereas that calculated for molecules within the AD was 0.80, which obviously did not differ from that calculated in the whole dataset (0.79) since the vast majority of the molecules were

included in the AD. Four of the five identified classes were predicted with a R^2 greater than that calculated for in AD molecules: hydrazones (146 molecules, R^2 0.87), sulfoxides (43 molecules, R^2 0.82), CRX3 (349 molecules, R^2 0.83) and imidazoles (252 molecules, R^2 0.85).

The T.E.S.T. software, which resulted as the second best model from the ANTARES evaluation (R^2 0.68), evaluated only molecules which fell within its AD. The R^2s calculated for the five identified classes were always higher than the global one: 0.83 for hydrazones, 0.76 for sulphides, 0.82 for sulfoxides, 0.78 for CRX3, and 0.78 for imidazoles. The performance for these classes were lower compared to ACD, however the increment with respect to the global R^2s were higher. T.E.S.T. seemed therefore to benefit more than ACD of the possible use these classes for the AD definition.

The TerraQSAR software did not include an AD evaluation tool. Four of the identified chemical classes were predicted with a R^2 higher than the global one (0.64): 0.71 for hydrazones, 0.89 for sulfoxides, 0.70 for CRX3, and 0.74 for imidazoles. Generally, the improvement of the predictive performance were comparable to those obtained for ACD. Sulfoxide molecules, however, gave the greatest improvement and also the greatest R^2 with respect to the other models. This class was however probably the less significant due to its low presence within the dataset (43 molecules out of 7417).

The global performance of the ADMET Predictor software were rather poor: 7293 molecules within its AD were predicted with a R^2 of 0.54, whereas the very few molecules excluded from the AD (123) were predicted with a R^2 of 0.17. The evaluation of the five identified classes resulted in a R^2 always greater than that calculated for

molecules within the AD: 0.71 for hydrazones, 0.63 for sulphides, 0.67 for sulfoxides, 0.70 for CRX3, and 0.74 for imidazoles.

Finally, TOPKAT gave very low performance for the dataset used, and the built-in reliability index (used for the AD definition) did not improved the situation: 6610 molecules within its AD were predicted with a R^2 of 0.41, whereas the other 807 molecules gave a R^2 of 0.28. The performance evaluated on the five classes, even if not very high (R^2 between 0.50 and 0.55), suggested that even in this case the chemical classes approach could improve the model AD definition.

With the aim of identifying common potentially problematic molecules, ten classes predicted with the lowest R^2 were also considered for each model. Seven classes were always predicted with a R^2 substantially lower compared to the whole dataset, suggesting their possible implementation as AD rules (Table 26 - Chapter F). The occurrences of these classes were however rather low, giving potentially the possibility to exclude only a small number of molecules from the models' applicability domain, and behaving, in fact similarly to the built-in AD evaluation tool.

ACD predicted the identified classes with the highest R^2, which ranges from 0.37 to 0.55, compared to the other models, confirming the global results obtained within the ANTARES project. Oxazoles (23 molecules - R^2 0.39), thioanhydrides (34 molecules - R^2 0.37) and aromatic aldehydes (36 molecules - R^2 0.40) were the worst predicted classes. However, their occurrence were too low to possibly provide a significant improvement to the model's AD. The class with the high number of occurrence was aromatic secondary amines: 174 molecules were predicted with a R^2 of 0.48. However more than

the molecules excluded using the model's built-in AD evaluation, the R^2 was also higher, suggesting that its application would not affect the model performance.

The R^2 values obtained for the T.E.S.T. software showed an opposite scenario. All the classes were predicted with a very low R^2, which ranges from 0.01 to 0.34. From the applicability domain point of view, these results were significant, especially considering that T.E.S.T. included all of them within its AD. Moreover, pyrroles and aromatic secondary amines, which were the classes with the highest occurrence (174 and 81 molecules), were predicted with very low R^2 (0.22 and 0.26), suggesting a possible impact for the AD definition.

The results for TerraQSAR were similar, apart from thioanhydrides that were predicted with a R^2 of 0.76 (greater compared to the global dataset). The R^2 calculated for the other classes ranged from 0.12 to 0.34. Also in this case pyrroles and aromatic secondary amines were predicted with very low R^2 (0.31 and 0.21). The results obtained suggested a possible use of the identified chemical classes also for this model.

ADMET predictor and TOPKAT resulted as the model potentially more affected by the use of the selected classes for the AD definition. The R^2 values ranged from 0.00 to 0.28 for ADMET and from 0.01 to 0.09 for TOKAT. This reflected the poor predictivity of these models, which were also highlighted within the ANTARES analysis. For both models the classes gave performance worse compared to the molecules excluded by the model's AD. TOPKAT gave a R^2 of 0.28 for the 807 molecules excluded from its AD, whereas ADMET Predictor gave a R^2 of 0.17 for 123 molecules. Also in this case, the results supported the chemical classes-based approach.

To conclude, the results suggested that the identified rules could improve the applicability domain definition for all the models considered. These classes, however, cannot be used *per se*, since they are potentially able to define the applicability domain only for molecules belonging to them. The integration of these rules within the available AD definitions would be the most preferable solution.

The good results obtained using the chemical classes-based approach, led to a more ambitious aim: an *a priori* definition of (Q)SAR models' applicability domain. The idea was to hypothesize which chemical classes could be better predicted for a certain endpoint, starting from end endpoint distribution among each class. A new endpoint was chosen as case study: mutagenicity. Also in this case, a large dataset (6064 molecules) of experimental Ames test values was available through the ANTARES project. The working hypothesis was that (Q)SAR models should be able to predict chemical classes whose molecules showed the same effect (e.g. near all mutagens) with a higher accuracy compared to more heterogeneous ones (e.g. classes composed by 50% mutagens and 50% non-mutagens). The preliminary analysis identified 32 chemical classes whose mutagenicity (defined in this case as the percentage of mutagenic molecules) substantially deviated from the global dataset (Figure 7 - Chapter F). The dataset used for mutagenicity was really balanced with regards of mutagenic and non-mutagenic molecules (54% were mutagens), a variation of +/-20% was used to considered chemical classes as relevant for the AD definition. A simple approach was initially used, considering the molecules of these chemical classes as within the *a priori* applicability domain. This resulted in considering the 58% of the dataset as not reliably

predictable by (Q)SAR models. This simple and preliminary definition was used to define the applicability domain of the three (Q)SAR models included in the freely available VEGA software: CAESAR, SARpy, and the Benigni-Bossa ruleset for mutagenicity. The reasons for this choice were related to several aspects: the modelling approaches used to build them were different, CAESAR was a hybrid model composed on a molecular descriptor-based model and two lists of structural alerts (derived from the Benigni-Bossa ruleset), SARpy was composed by structural fragments statistically related to the mutagenicity effect (developed using the SARpy software), and the Benigni-Bossa was a knowledge-based model, composed by structural alerts experimentally related to mutagenicity effects (this ruleset has been also used to develop the commercial software ToxTree). Moreover, all the models and the VEGA platform have been developed within our research group, therefore the exploitation of the results obtained should be easier.

The use of the chemical classes-based applicability domain definition was evaluated considering information about the training sets used to build the models. Generally, (Q)SAR models predict the molecules used to build them with higher accuracy compared to "new" chemicals. The presence of such molecules within the dataset used for the study could therefore introduce a bias in the analysis. VEGA automatically evaluates the applicability domain and identifies molecules that fall within it, the evaluation performed by VEGA was therefore compared to that obtained using the chemical classes. Considering the subset of molecules not in common with the model's training sets (3026 molecules), the models performance slightly improved.

The CAESAR model predicted the external molecules with high accuracy (0.73) and sensitivity (0.85). Using the chemical classes both these parameters slightly increased for molecules within the AD (1299 molecules, accuracy 0.77 and sensitivity 0.87) and decreased for the other molecules (1727 molecules, accuracy 0.71 and sensitivity 0.83), however the difference between in AD and out AD was too small, showing that chemical classes were not able to correctly discriminate between reliable and unreliable predictions. The specificity parameters were also characterized by similar variations (global 0.60, in AD 0.62 and out AD 0.59). The VEGA built-in AD tool, on the contrary, substantially improved the performance for new molecules, predicting those within the AD with an accuracy of 0.83, a sensitivity of 0.93, and a specificity of 0.71. Moreover, about one thousand more molecules were included within the AD, compared to the chemical classes approach. The performance also differed substantially compared to molecules outside the AD (accuracy 0.49, sensitivity 0.61, and specificity 0.40). The analysis of both SARpy and the Benigni-Bossa ruleset showed more or less the same scenario: chemical classes only slight improved the performance for new molecules which fell within the AD, but left a high number of reliable predictions outside the applicability domain.

All the models predicted a number of the selected chemical classes with low accuracy. In the cases of CAESAR and SARpy, only a few classes were badly predicted, and were also generally composed of few molecules. Benigni-Bossa predicted a higher number of classes with low accuracy, including also classes composed of hundreds of molecules. The impact of these classes on the AD definition could be a target for future

analysis. On the other hand, among the selected classes, nitroaromatic compounds could have had a high impact on the definition of the AD. This functional group is well-known related to mutagenic effect, indeed is included among the Benigni-Bossa ruleset, as a structural alert for mutagenicity. This class was one of the most common among the whole dataset, and was also the most numerous among the selected classes. The models were generally able to correctly predict these molecules, therefore the presence of this class could have played a prominent role in the AD definition. This could be another possible target for future studies.

To conclude, considering especially the CAESAR and SARpy model, the possible integration of the identified chemical classes within the built-in VEGA AD tool could still improve the model performance. As already explained, VEGA uses several parameters to calculate the so-called Applicability Domain Index (ADI), which ranges from 0 (unreliable prediction) to 1 (reliable prediction). The presence of the functional groups used to define the identified chemical classes could be used as a further parameter for the ADI calculation. In particular, since the performance increased for molecules composed by these groups, their presence could be used to increase the value of the ADI.

Coming back to the subset of classes badly predicted by the model, a possible reason was thought to be related to the techniques used to build the (Q)SAR models analysed. If the hypothesized relation between "purest" classes and models' predictability was correct, a trend should have been identified by plotting the percentage of mutagens against the prediction accuracy. In particular, the chemical classes should have formed

a "U" or a "V" shape. The CAESAR models seemed more "adherent" to the hypothesis than SARpy and the Benigni-Bossa ruleset (Figure 6 - Chapter F). Interestingly, while the three models were based on structural alerts, only CAESAR included also chemical descriptors. The further analysis of three chemical classes chosen as case studies supported this explanation.

The nitroaromatic functional group was found in 898 molecules, and 85% of them were mutagens. Moreover, the three models gave high accuracy for this class: 0.86 for both CAESAR and Benigni-Bossa, and 0.83 for SARpy. As already explained, the nitroaromatic functional group is well-known to be related to mutagenic effect, and is a structural alert present in both Benigni-Bossa and CAESAR. Its presence is sufficient, for these two models, to predict a molecule as mutagen. However, some exceptions were described by Benigni-Bossa. This is the case of carboxylic acid and a sulfonic acid group. Nitroaromatic molecules bearing one of these two groups are predicted as non-mutagens. The influence of secondary classes of the prediction accuracy was therefore investigated, resulting in a nearly linear correlation between the mutagenicity of these classes (again expressed as the percentage of mutagens) and the prediction accuracy of both CAESAR and Benigni-Bossa. Only four outliers were identified, two of them were the sulfonic and the carboxylic acids. The number molecules of molecules within these secondary classes were too few to be further investigated. Also SARpy showed a correlation, however less strong and with more outliers. This model is not based on functional groups but on structural fragments. The analysis of the presence of the

nitroaromatic functional group (and its exceptions) within its fragments could help in better understand this, even less clear, correlation.

The same analysis was performed for another highly mutagenic chemical class: aliphatic hydroxylamines. In this case the Benigni-Bossa ruleset and the CAESAR model do not list the functional group as a structural alert. The scatter plot analysis showed again a linear trend between the mutagenicity of the secondary classes and the accuracy in prediction. In this case the accuracy values were less aligned compared to nitroaromatic molecules. However the trend was clear also for SARpy: the accuracy decreased for classes richer in mutagens. The comparison of the scatter plots showed that the trend for CAESAR model was "shifted" towards better accuracy, compared to Benigni-Bossa. Since no structural alert is present for aliphatic hydroxylamines, both models failed to predict secondary classes mainly composed of mutagenic molecules. The trend showed by the SARpy model supported this idea. For most of the secondary classes, the trend was quite similar (in terms of accuracy) to that showed by Benigni-Bossa. Instead, all the classes composed by 60 to 80% of mutagens gave accuracy comparable to CAESAR. Both the CAESAR molecular descriptor module and the SARpy model were built starting from the same training set. Most probably, both these models were able to partially learn how to predict this chemical class during their training. A further analysis should be to check the composition of the CAESAR/SARpy training set to verify this hypothesis.

The last class considered as case study was aliphatic tertiary amides. This class was identified as one of the possibly less mutagen (only 25% of the molecules were

mutagens). The Benigni-Bossa ruleset and VEGA do not include rules for non-mutagens, apart from the exceptions to the structural alert, however this the aliphatic tertiary amides functional group was not listed as exception to any rule. The scatter plot analysis of the secondary classes showed a trend also for these molecules. In this case both CAESAR and Benigni-Bossa seemed able to better predict less mutagenic secondary classes. The SARpy model, on the other hand, did not show a particular trend.

In conclusion, the secondary classes approach seems to be able to provide interesting *a priori* information for the definition of SAR models' applicability domain. Better results should be obtained for secondary classes mainly composed of mutagenic molecules, if the functional group that defines the primary class is also a structural alert for the model. This simple observation could be helpful in the improvement of the applicability domain of each structural alert present in a ruleset. On the other hand, if the primary class is not defined by a functional group present among the model's structural alerts, and is mainly composed by mutagens, better performance should be expected for less mutagenic secondary classes. Finally, the opposite behaviour is expected for primary classes not present among the model's rules and composed mainly by non-mutagenic molecules. Interestingly, if we consider the scatter plot obtained in this case together with that obtained in the first case, the expected "U" shape can be observed.

Starting from the results obtained from analysis if the "secondary classes" approach, a tool is currently being developed within our group, able to identify primary and secondary rules, automatically. The best version of this software was used for a

preliminary analysis. Primary and secondary classes which showed a substantial difference (in terms of either mutagenicity or accuracy) were extracted using this software. Also this tool identified the aliphatic hydroxylamines class as potentially relevant, since its mutagenicity was substantially higher compared to the global dataset. This class was also identified as related to a decrease in performance for the three models considered. Interestingly, secondary classes which furtherly substantially decreased the accuracy were identified for the CAESAR and SARpy models. In the first case the classes were all related to the presence of aromatic rings, whereas in the second case the number of nitrogen atoms seemed to negatively affect the SARpy performance for aliphatic hydroxylamines. The analysis performed was simple and preliminary, however the results supported the secondary classes-based approach for the determination of the (Q)SAR models applicability domain and showed the potential of the automation of this approach.

The SARpy software is a freely available tool able to identify structural fragments related to a binary property chosen by the user. It has been used for example to develop the SARpy model for mutagenicity, included in the VEGA platform. Compared to the functional groups and atom-centered used by istChemFeat, these structural fragments are generally more complex and therefore potentially more specific. The idea was to use this software to model the errors in prediction of (Q)SAR models, by extracting structural features related to correct or wrong predictions. The CAESAR model for mutagenicity was chosen as case study. This software can classify molecules as "Mutagen", "Suspect Mutagen" or "Non Mutagen", however a binary classification was adopted within the

ANTARES project, by considering the "Suspect Mutagen" as "Mutagen", accordingly to the conservative approach foreseen within regulatory contexts. Therefore, two type of errors can be identified: mutagenic molecules predicted as non-mutagens (false negative prediction, FN), or non-mutagenic molecules predicted as mutagens (false positive prediction, FP). Molecules correctly predicted are referred as true positive (TP) if mutagens, and true negative (TN) if non-mutagens. Two rulesets were extracted with SARpy, to consider the two type of errors separately. The first ruleset included fragments related to TP and FP CAESAR's predictions, whereas the second one included fragments related to TN and FN. This resulted in the identification of 125 fragments related to TP predictions, 40 for FP, 78 for TN, and 18 for FN. The number of fragments related to correct predictions were higher because CAESAR produced a relatively small number of errors (813 FP and 295 FN) compared to the whole dataset (6065 molecules), therefore SARpy had only few molecules to study. The prediction accuracy of these rulesets were evaluated using test sets composed by new molecules (not used for their extraction), and the results supported the use of these fragments for the estimation of correct and wrong CAESAR's predictions (Table 13 and Table 15 - Chapter F). Both TP-FP and TN-FN rulesets predicted wrong and correct predictions with an accuracy of 0.74. Being based on structural fragments, these rulesets could predict only molecules containing them. Therefore, TP-FP and TN-FN could not predict 15% and 31% of their test sets.

What we obtained was substantially a four-rulesets-based model for the prediction of correct (TP and TN) and wrong (FP and FN) CAESAR's predictions. These rulesets were

used to define the applicability domain of the CAESAR model, by considering molecules predicted as TP and TN as within the AD, and those predicted as FP and FN as outside the AD. Molecules not predicted by the SARpy rulesets were left to be assessed by the built-in VEGA AD tool.

The rulesets-based and VEGA AD definitions were compared on the basis of the improvement they provide to the CAESAR performance. This comparison followed a four-step process: definition of the AD using the complete lists of fragments, study of the possible performance improvement by considering only the most "accurate" ones, analysis of the performance of the VEGA AD considering only the molecules for which the SARpy fragments could not provide AD info, and integration of the two AD definition. Moreover, the "worst" possible situation was considered, by analysing the CAESAR performance on completely new chemicals. This subset was obtained excluding molecules in common with the whole CAESAR dataset, not just its training set. Indeed, VEGA evaluates the AD using all the data available for the model, and compares the target molecule with those most similar present within the whole dataset. Therefore the presence of the target chemical within the dataset could simplify its evaluation. For the same reason, the molecules used as training set to identify the SARpy rules were also removed. The external validation set was finally composed of 762 molecules (10% of the original dataset). Considering the prediction accuracy, the two AD approaches gave comparable results: 0.75 and 0.76 for molecules within the ruleset-based AD and VEGA AD (Chapter F.19 - Table 16 and Table 17), and 0.53 and 0.56 for molecules without the ADs. The detailed analysis of FP and FN, however, highlighted some differences. Both

approaches were able to discriminate between mutagens correctly and wrongly predicted (TP and FN), indeed the models' sensitivity (0.89 for ruleset AD and 0.91 for VEGA AD) slightly increased compared to the whole external set (0.84). However, a lot of true positive predictions were excluded from the model's AD, especially by the rulesets: CAESAR's sensitivity for out AD molecules was 0.74. The VEGA AD performed better, with a sensitivity for out AD of 0.67, which was however still high. Considering the FP errors, the ruleset performed better in discriminating unreliable predictions. Both methods slightly increased the CAESAR specificity for in AD molecules (0.57 using the ruleset and 0.59 using VEGA) compared to the whole external set (0.55), suggesting that a lot of FP predictions were considered as reliable. The most interesting results were obtained, however, for molecules excluded from the applicability domain. Using the ruleset approach resulted in a specificity of 0.34, suggesting a subset mainly composed of FP predictions. The VEGA AD performed worse, providing a specificity of 0.47. These results opened the possibility of an integration of these two methods.

Before trying the VEGA-ruleset integration, however, a possible improvement of the ruleset's performance was studied. For each rule SARpy provides a score (called the likelihood ratio index, LRI) that substantially describes their prediction ability. Different LRI thresholds were used to verify the improvement of the CAESAR performance on the training set used to build ruleset (Chapter F.19 - Table 18). A reduced version of the original ruleset was obtained and evaluated on the SARpy test set (Chapter F.19 - Table 19). The reduced ruleset was not able to provide information for a higher number of molecules (921) compared to the original one (414). However, its ability to correctly

identify FP greatly increased: the specificity for out AD molecules decreased from 0.51 (obtained using the complete ruleset) to 0.15. A slight improvement was also observed for in AD molecules (the specificity increased from 0.74 to 0.79) and for sensitivity for out AD molecules (the sensitivity decreased from 0.85 to 0.77).

Another analysis was important before integrating the two approaches: checking the ability of VEGA to discriminate reliable and unreliable predictions for the molecules not considered by the SARpy ruleset (Chapter F.19 - Table 20 and Table 21). The differences between accuracy, sensitivity and specificity calculated for in AD and out AD molecules were used for the comparison. Considering again the new molecules, VEGA was more able to discriminate between reliable and unreliable predictions for those not considered by SARpy (accuracy, sensitivity and specificity differences between in AD and out AD were 0.24, 0.29 and 0.15), than on the whole subset (0.20, 0.24 and 0.12).

The final step of the study was the integration of the two methods. A very simple approach was used: the SARpy reduced ruleset was used to determine the AD information for the molecules containing its fragments, and VEGA was used for the other molecules (Chapter F.19 - Table 22). Even this simple approach produced interesting and promising results. The performance for in AD molecules were comparable to that obtained using VEGA. However, a far greater number of molecules were included within the AD (624, 82% of the external subset), compared to VEGA (524 – 69%). Even more important results were obtained considering molecules outside the applicability domain: accuracy, sensitivity and specificity resulted lower for the combined approach (0.45, 0.58 and 0.36) compared to VEGA (0.56, 0.67 and 0.47). Considering that the

number of molecules excluded from the combined approach (138) were substantially lower than those excluded by VEGA (238), it is possible to conclude that the introduction of the structural fragments information improved the definition of the CAESAR model applicability domain.

Chapter H.
Conclusions and future perspective

In the last years, the European Community funded several international research project with the aim of improving the knowledge and the use of computer-based predictive models. (Quantitative) Structure-Activity Relationship ((Q)SAR) represents a family of approaches commonly used to develop such models. These methods are based on the assumption that biological properties (such as toxicity) are strictly related to the structural conformation of chemicals. Therefore, using datasets of chemicals with known structure and experimental values of the property to analyse, it is possible to study the structure-activity relationship, and to also apply the "rules" extracted on new chemicals, to predict their properties. (Q)SAR methods are based on mathematical and statistical modelling approaches, which learn from a training set of experimental observations and can be then used to obtain predictions. Depending on the similarity of the new element compared to the training set, the estimation could be considered as an interpolation or an extrapolation. In the first case the results are usually more reliable. The problem of prediction reliability, therefore, affects also (Q)SAR models, making it indispensable to correctly identify the so-called "applicability domain" (AD) of the models.

As reported in Chapter C, several methods for the determination of the AD have been developed through years, including approaches based on chemical classes and atom-

centered fragments. In a recent study, the LIFE+ ANTARES project evaluated more than 50 (Q)SAR models for eight endpoints relevant for the REACH regulation. The performance of these models were assessed using large and verified dataset and considering, when available, the information about the AD of models, provided by their developers. The results obtained showed that the use of AD information generally improved the model's predictive performance and could be helpful in the discrimination of reliable and unreliable predictions.

The main aim of the research activities described in this thesis, was the study of the possibility to use structural parameters to improve the definition of (Q)SAR models' AD. The leading idea was to use structural properties and features of different complexity, from simple properties (e.g. molecular weight) to statistically extract molecular fragments. Different type of endpoints and models were also considered to obtain an overview of the possible application of the methods studied.

The first and simplest approach studied was based on the hypothesis that simple properties (e.g. molecular weight and chemical composition) could affect the reliability of model's prediction. The idea was to identify thresholds for these properties, able to discriminate between molecules within or outside the applicability domain of (Q)SAR models. This approach was studied using the CAESAR model for bioconcentration factor, which was a rather simple neural network, based on eight molecular descriptors. The results showed that threshold selected using a training set were also able to improve the definition of the applicability domain of the CAESAR model while applied on new molecules.

These results opened the way for more complex analysis that involved the use of functional groups and structural fragments. Functional groups were used to define chemical classes, whose possible relation with model's reliability was then investigated. The leading hypothesis was to identify chemical classes either commonly correctly or wrongly predicted by different models for the same endpoint. A preliminary study of five models for oral rat acute toxicity confirmed this hypothesis, leading to the identification of five chemical classes predicted with high R^2 values and seven with low R^2s by all the models considered. These results seemed to support the idea of using chemical classes for an *a priori* definition of (Q)SAR models' applicability domain. This possibility was investigated using three models for mutagenicity. The change in the endpoint was mainly due to the possibility of integrating the results obtained, within the VEGA platform (developed within our research group), which did not include models for acute toxicity. To define the AD *a priori*, it was hypothesized that chemical classes commonly composed of either mutagens or non-mutagens, should by easier to predict by models. The results initially obtained only partially supported this hypothesis. Starting from the experimental evidences, included within the Benigni-Bossa ruleset for mutagenicity, that the contemporary presence of a secondary functional group could decrease the mutagenicity of a primary one, the influence of secondary classes on the accuracy in prediction was also investigated. The results suggested that the use of secondary classes could help in improving the applicability domain definition, however more studies are necessary to confirm generalize this hypothesis.

Finally, the possibility of modelling the errors in predictions using a structural alerts-based approach was studied. Structural alerts can be generally used to build SAR models (e.g. for mutagenicity), and can be derived by experimental evidences or using statistical methods. In this case, the prediction correctness of the CAESAR model for mutagenicity was used as an endpoint to identify statistically relevant fragments. Fragments related to correct predictions were used to identify molecules within CAESAR's AD, whereas those related to wrong predictions were used to exclude molecules from the AD. The obtained ruleset was compared and integrated with the AD definition provided by CAESAR, proving its ability to improve this definition.

All the approaches presented in this study are of course open for further analysis. New chemicals and endpoints, as well as other available models could be used to study the reliability of the methods developed. The possible overlap between the approaches presented in this thesis and those provided by the models' developers should also be studied more in detail. In order to further confirm the improvements obtained within this work, the molecules classified as within the model AD (as well as those excluded from it) by different approaches should be compared. Moreover, the integration of the approaches presented could help in improving even more the definition of the applicability domain of (Q)SAR models. Since the methods developed are all based on structural features (atom, functional groups and structural fragments), possible overlaps could be identified while applying them to the same endpoint and models. For example, the functional groups used for the definition of chemical classes, could be also present within the structural alerts statistically identified. Such results would confirm the

importance of certain features for the AD definition, and the possibility to identify chemicals easily predicted, for a particular endpoint, by virtually every model developed. The opposite situation would be an even more important goal: identifying classes of molecules difficult to be reliably predicted could on one hand stimulate the research of new modelling approaches, and other hand could suggest the end-users to use other methods to obtain the data needed. More realistically, an improved definition of (Q)SAR models' AD is of key importance for the integration of the results obtained by different models, for example in the case of risk assessment.

References

1. American Chemical Society's Chemical Abstract Service (CAS) - http://www.cas.org/index
2. Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) – Annex XI: "GENERAL RULES FOR ADAPTATION OF THE STANDARD TESTING REGIME SET OUT IN ANNEXES VII TO X". European Parliament and Council Regulation (EC) No 1907/2006.
3. Johnson M, Maggiora GM, Eds. Concepts and Applications of Molecular Similarity; John Wiley & Sons: New York, 1990.
4. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? J Med Chem. 2002 Sep 12;45(19):4350-8.
5. Sheridan RP. Finding multiactivity substructures by mining databases of drug-like compounds. J Chem Inf Comput Sci. 2003 May-Jun;43(3):1037-50.
6. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug Discov Today. 2002 Sep 1;7(17):903-11.
7. Schneider G, Böhm HJ. Virtual screening and fast automated docking methods. Drug Discov Today. 2002 Jan 1;7(1):64-70.
8. Schuffenhauer A, Gillet VJ, Willett P. Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. J Chem Inf Comput Sci. 2000 Mar;40(2):295-307.
9. Borman S. New QSAR Techniques Eyed for Environmental Assessments. Chem. Eng. News, 1990;68;20-23.

10. Lipnick RL. Charles Ernest Overton; Narcosis Studies and a Contribution to General Pharmacology. Trends Pharmacol. Sci. 1986;7:161-4.

11. Hansch C, Fujita T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Am Chem Soc. 1964;86(8):1616-26.

12. Unger SH, Hansch C. On model building in structure-activity relationships. A reexamination of adrenergic blocking activity of beta-halo-beta-arylalkylamines. J Med Chem. 1973 Jul;16(7):745-9.

13. Lombardo A, Schifanella O, Roncaglioni A, Benfenati E. Quantitative Structure-Activity Relationship (QSAR) in Ecotoxicology. In: Férard J-F, Blaise C (Ed.). "Encyclopedia of Aquatic Ecotoxicology". Springer Netherlands (2013), 945-056.

14. Benigni R, Bossa C, Jeliazkova N, Netzeva T, Worth A. The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree. JRC Scientific and Technical Reports. 2008.

15. Dimitrov S, Dimitrova N, Parkerton T, Comber M, Bonnell M, Mekenyan O. Base-line model for identifying the bioaccumulation potential of chemicals. SAR QSAR Environ Res. 2005 Dec;16(6):531-54.

16. Piegorsch WW, Zeiger E. Measuring intra-assay agreement for the Ames salmonella assay. In: Hotorn L (Ed.). "Statistical Methods in Toxicology, Lecture Notes in Medical Informatics.". Springer-Verlag (1991), 35–41.

17. Benfenati E, Boriani E, Craciun M, Malazizi L, Neagu D, Roncaglioni A. Databases for pesticide ecotoxicity. In: Benfenati E. (Ed.). "Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes", Elsevier Science Ltd, Amsterdam, The Netherlands (2007), 59-81.

18. Zhao C, Boriani E, Chana A, Roncaglioni A, Benfenati E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). Chemosphere. 2008 Dec;73(11):1701-7.

19. DRAGON, Talete S.R.L.

20. ADMET Predictor. Simulations Plus Inc., Lancaster, USA; software available at http://www.simulations-plus.com/

21. Katritzky AR, Karelson M, Petrukhin R. COmprehensive DEscriptors for Structural and Statistical Analysis (CODESSA); software available at http://www.codessa-pro.com/

22. Ferrari T, Gini GC, Bakhtyari NG, Benfenati E. Mining toxicity structural alerts from SMILES: A new way to derive Structure Activity Relationships. In: CIDM(2011)120-127

23. Todeschini R, Consonni V. Handbook of Molecular Descriptors. John Wiley & Sons (2008)

24. The IUPAC International Chemical Identifier (InChI) - http://www.iupac.org/home/publications/e-resources/inchi.html

25. Daylight SMILES (Simplified Molecular Input Line Entry System) - http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html

26. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J. Chem. Inf. Comput. Sci. 1992;32(3):244-255.

27. Devillers J. (Ed.). Genetic Algorithms in Molecular Modeling. Principles of QSAR and Drug Design, Elsevier Science Ltd, Amsterdam, The Netherlands (1996), 327 pp.

28. Devillers J. (Ed.). Neural Networks in QSAR and Drug Design. Principles of QSAR and Drug Design, Elsevier Science Ltd, Amsterdam, The Netherlands (1996), 284 pp.

29. Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model. 2010 Jul 26;50(7):1189-204.

30. Varmuza K, Filzmoser P. Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton, FL, USA (2009).

31. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004 Jan-Feb;44(1):1-12.

32. Livingstone DJ, Rahr E. Corchop – an Interactive Routine for the Dimension Reduction of Large QSAR Data Sets. Quant. Struct.-Act. Relat. 1989;8(2):103-108.

33. Whitley DC, Ford MG, Livingstone DJ. Unsupervised forward selection: a method for eliminating redundant variables. J Chem Inf Comput Sci. 2000 Sep-Oct;40(5):1160-8.

34. Frimayanti N, Yam ML, Lee HB, Othman R, Zain SM, Rahman NA. Validation of quantitative structure-activity relationship (QSAR) model for photosensitizer activity prediction. Int J Mol Sci. 2011;12(12):8626-44.

35. Purohit V, Basu AK. Mutagenicity of nitroaromatic compounds. Chem Res Toxicol. 2000 Aug;13(8):673-92.

36. RS, Wang C, Hughes JB, Kutty R, Bennett GN. Mutagenicity of nitroaromatic degradation compounds. Environ Toxicol Chem. 2003 Oct;22(10):2293-7.

37. OECD. (2004). Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. Paris, France: OECD.

38. Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK. Validation of QSAR Models - Strategies and Importance. Int J of Drug Disc. 2011 Jul-Sep;2(3): 511-519

39. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inf. 2010;29(6-7):476-88.

40. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. J Med Chem. 2014 Apr 24;57(8):3186-204.

41. Maggiora GM. On outliers and activity cliffs--why QSAR often disappoints. J Chem Inf Model. 2006 Jul-Aug;46(4):1535.

42. Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. J Med Chem. 2012 Apr 12;55(7):2932-42.

43. Stumpfe D, Hu Y, Dimova D, Bajorath J. Recent progress in understanding activity cliffs and their utility in medicinal chemistry. J Med Chem. 2014 Jan 9;57(1):18-28.

44. Brown RD, Martin YC. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. J. Chem. Inf. Comput. Sci. 1997 Jan 1;37(1):1-9

45. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon J-F, Cornell WD. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. J. Chem. Inf. Model. 2007 Jun 26;47(4):1504–19.

46. Fliri A, Loging W, Thadeio PF, Volkmann R. Biological Spectra Analysis: Linking Biological Activity Profiles to Molecular Structure. Proc. Natl. Acad. Sci. U.S.A. 2005 Jan 11;102(2):261–266.

47. Petrone PM, Simms B, Nigsch F, Lounkine E, Kutchukian P, Cornett A, Deng Z, Davies JW, Jenkins JL, Glick M. Rethinking molecular similarity: comparing compounds on the basis of biological activity. ACS Chem Biol. 2012 Aug 17;7(8):1399-409.

48. Mason JS, Good AC, Martin EJ. 3-D Pharmacophores in Drug Discovery. Curr. Pharm. Des. 2001;7(7):567–597.

49. Renner S, Schneider G. Scaffold-hopping potential of ligand-based similarity concepts. ChemMedChem. 2006 Feb;1(2):181-5.

50. Brooks DG, Carroll SS, Verdini WA. Characterizing the Domain of a Regression Model. The American Statistician. 1988 Aug;42(3):187-190

51. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MT, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts D, Schultz T, Stanton DW, van de Sandt JJ, Tong W, Veith G, Yang C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim. 2005 Apr;33(2):155-73.

52. Veith GD. On the nature, evolution and future of quantitative structure-activity relationships (QSAR) in toxicology. SAR QSAR Environ Res. 2004 Oct-Dec;15(5-6):323-30.

53. Schultz TW, Cronin MT. Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships. Environ Toxicol Chem. 2003 Mar;22(3):599-607.

54. Cefic (2002). (Q)SARs for Human Health and the Environment. In Workshop Report on Regulatory acceptance of (Q)SARs Setubal, Portugal, March 4-6, 2002

55. Alternative Non-Testing methods Assessed for REACH Substances (ANTARES). http://www.antares-life.eu/

56. Pizzo F, Lombardo A, Manganaro A, Benfenati E. In silico models for predicting ready biodegradability under REACH: a comparative study. Sci Total Environ. 2013 Oct 1;463-464:161-8.

57. Cappelli CI, Manganelli S, Lombardo A, Gissi A, Benfenati E. Validation of quantitative structure-activity relationship models to predict water-solubility of organic compounds. Sci Total Environ. 2013 Oct 1;463-464:781-9.

58. Gissi A, Nicolotti O, Carotti A, Gadaleta D, Lombardo A, Benfenati E. Integration of QSAR models for bioconcentration suitable for REACH. Sci Total Environ. 2013 Jul 1;456-457:325-32.

59. Bakhtyari NG, Raitano G, Benfenati E, Martin T, Young D. Comparison of in silico models for prediction of mutagenicity. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev. 2013;31(1):45-66.

60. Gonella Diaza R, Manganelli S, Esposito A, Roncaglioni A, Manganaro A, Benfenati E. Comparison of in silico tools for evaluating rat oral acute toxicity. SAR QSAR Environ Res. 2015 Jan;26(1):1-27

61. AMBIT Discovery. Ideaconsult Ltd., Sofia, Bulgaria, 2008; ; software available at http://ambit.sourceforge.net/

62. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Altern Lab Anim. 2005 Oct;33(5):445-59.

63. Tubbs JD. A note on binary template matching. Pattern Recognit. 1989 Aug;22(4):359-365.

64. Barratt MD, Basketter DA, Chamberlain M, Admans GD, Langowski JJ. An expert system rulebase for identifying contact allergens. Toxicol In Vitro. 1994 Oct;8(5):1053-60.

65. TOPKAT OPS (2000) , US patent no. 6 036 349 issued March 14, 2000

66. Preparata FP, Shamos MI. Convex Hulls: Basic Algorithms, In: "Computational Geometry: An Introduction". Springer-Verlag, New-York (1991), 95-148.

67. Jaworska J, Aldenberg T, Nikolova N, Review of methods for assessing the applicability domains of SARs and QSARs. Final report to the Joint Research Centre (Contract No. ECVA-CCR. 496575-Z). Part 1: Review of statistical methods for QSAR AD estimation by the training set. 2005.

68. Stanton DT, Jurs PC, Hicks MG. Computer-assisted prediction of normal boiling points of furans, tetrahydrofurans, and thiophenes. J. Chem. Inf. Comput. Sci. 1991 May 1;31(2):301-310.

69. Stanton DT, Egolf LM, Jurs PC, Hicks MG. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. J. Chem. Inf. Comput. Sci. 1992 Jul 1;32(4):306-316

70. Stanton DT. Development of a Quantitative Structure–Property Relationship Model for Estimating Normal Boiling Points of Small Multifunctional Organic Molecules. J. Chem. Inf. Comput. Sci. 2004 Jan 1;40(1):81-90

71. Tropsha A, Gramatica P, Gombar V. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR Models. QSAR Comb. Sci. 2003 Apr;22(1):69-77.

72. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect. 2003 Aug;111(10):1361-75.

73. Neter J, Kutner MH, Wasserman W, Nachtsheim C. (1996). Applied Linear Statistical Models. 1408pp. New York, NY, USA: McGraw-Hill.

74. Fukunaga K. (1990). Introduction to Statistical Pattern Recognition. 2nd edn, 592pp. Computer Science and Scientific Computing Series. New York, NY, USA: Academic Press.

75. Myers, R.H. (2000). Classical and Modern Regression with Applications. 2nd edn, 488pp. Pacific Grove, CA, USA: Duxbury Press.

76. Silverman BW. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London (1986), 176pp.

77. Gray A, Moore A. Nonparametric Density Estimation: Toward Computational Tractability. In: "Proceedings of SIAM International Conference on Data Mining 2003", San Francisco, USA (2003).

78. Gray A, Moore A. Very fast multivariate kernel density estimation using via computational geometry. In: "Proceedings of Joint Statistics Meeting 2003", The American Statistical Association, Alexandria, VA, USA (2003).

79. Rogers DJ, Tanimoto TT. A Computer Program for Classifying Plants. Science. 1960 Oct 21;132(3434):1115-8.

80. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD, Assessing Applicability Domains of Toxicological QSARs: Definition, Confidence in Predicted Values, and the Role of Mechanisms of Action. QSAR Comb. Sci. 2007 Feb;26(2):238-254.

81. Dragos H, Gilles M, Alexandre V. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. J Chem Inf Model. 2009 Jul;49(7):1762-76.

82. Virtual models for property Evaluation of chemicals within a Global Architecture (VEGA) – http://www.vega-qsar.eu/

83. Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39.

84. DAYLIGHT: Fingerprints - Screening and Similarity - http://www.daylight.com/dayhtml/doc/theory/theory.finger.html

85. Guha R, Dutta D, Jurs PC, Chen T. Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions. J. Chem. Inf. Model. 2006 May 19;46(4):1836–1847.

86. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. J Chem Inf Model. 2006 Sep-Oct;46(5):1984-95.

87. Feher M, Ewing T. Global or local QSAR: Is there a way out? QSAR Comb. Sci. 2009 Apr 29;28(8):850–855.

88. Helgee EA, Carlsson L, Boyer S, Norinder U. Evaluation of quantitative structure-activity relationship modeling strategies: local and global models. J Chem Inf Model. 2010 Apr 26;50(4):677-89.

89. Wood DJ, Buttar D, Cumming JG, Davis AM, Norinder U, Rodgers SL. Automated QSAR with a hierarchy of global and local models. Mol. Inf. 2011 Nov 15;30(11-12):960–972

90. Buchwald F, Girschick T, Seeland M, Kramer S. Using local models to improve (Q)SAR predictivity. Mol. Inf. 2011 Mar 14;30(2-3):205–218.

91. Davis AM, Wood DJ. Quantitative structure-activity relationship models that stand the test of time. Mol. Pharmaceutics 2013 Jan 14;10(4):1183–1190.

92. Zhang H, Ando HY, Chen L, Lee PH. On-the-fly selection of a training set for aqueous solubility prediction. Mol Pharm. 2007 Jul-Aug;4(4):489-97.

93. Sommer S, Kramer S. Three data mining techniques to improve lazy structure-activity relationships for noncongeneric compounds. J Chem Inf Model. 2007 Nov-Dec;47(6):2035-43.

94. Ellison CM, Sherhod R, Cronin MT, Enoch SJ, Madden JC, Judson PN. Assessment of methods to define the applicability domain of structural alert models. J Chem Inf Model. 2011 May 23;51(5):975-85.

95. Kühne R, Ebert R-U, Schüürmann G. Prediction of the temperature dependency of Henry's law constant from chemical structure. Environ Sci Technol. 2005 Sep 1;39(17):6705-11.

96. Dimitrov SD, Low LK, Patlewicz GY, Kern PS, Dimitrova GD, Comber MH, Phillips RD, Niemela J, Bailey PT, Mekenyan OG. Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. Int J Toxicol. 2005 Jul-Aug;24(4):189-204.

97. Ellison CM, Enoch SJ, Cronin MT, Madden JC, Judson P. Definition of the applicability domains of knowledge-based predictive toxicology expert systems

by using a structural fragment-based approach. Altern Lab Anim. 2009 Nov;37(5):533-45.

98. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. J. Chem. Inf. Comput. Sci. 1985 May 1;25(2):64-73.

99. Toxmatch; Ideaconsult Ltd.: Sofia, Bulgaria, 2008.

100. Kazius J, McGuire R, Bursi R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. J. Med. Chem. 2005 Jan 1;48(1):312-320

101. Kühne R, Ebert R-U, Schüürmann G. Chemical domain of QSAR models from atom-centered fragments. J Chem Inf Model. 2009 Dec;49(12):2660-9.

102. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan O. A stepwise approach for defining the applicability domain of SAR and QSAR models. J Chem Inf Model. 2005 Jul-Aug;45(4):839-49.

103. Kühne R, Kleint F, Ebert R-U; Schüürmann G. In: Gasteiger J (Ed.) "Software Development in Chemistry 10". PROserv Springer Produktionsgesellschaft, Berlin, Germany, pp 125-134.

104. Kühne R, Ebert R-U, Schüürmann G. Model selection based on structural similarity-method description and application to water solubility prediction. J Chem Inf Model. 2006 Mar-Apr;46(2):636-41.

105. Kühne R, Ebert R-U, Schüürmann G. Estimation of Compartmental Half-Lives of Organic Compounds - Structural Similarity Versus EPISuite. QSAR Comb. Sci. 2007 Apr;26(4):542–549.

106. Kühne R, Ebert R-U, Kleint F, Schmidt G, Schüürmann G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. Chemosphere. 1995 Jun;30(11):2061-77.

107. Abraham MH, Andonian-Haftvan J, Whiting GS, Leo A, Taft RS. Hydrogen Bonding. Part 34. The Factors That Influence the Solubility of Gases and Vapors in Water at 298 K, and a New Method for Its Determination. J. Chem. Soc., Perkin Trans. 2 1994, 1777-91.

108. von der Ohe PC, Kühne R, Ebert R-U, Altenburger R, Liess M, Schüürmann G. Structural alerts--a new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. Chem Res Toxicol. 2005 Mar;18(3):536-55.

109. REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC.

110. Pence HE, Williams A. ChemSpider: An Online Chemical Information Resource. J. Chem. Educ. 2010 Aug 30;87(11):1123-4

111. ChemSpider Home Page. http://www.chemspider.com/

112. Tomasulo P. ChemIDplus-Super Source for Chemical and Drug Information. Med Ref Serv Q. 2001 Jan 1;21(1):53-59

113. ChemIDplus Home Page. http://chem.sis.nlm.nih.gov/chemidplus/

114. European Centre for Ecotoxicology and Toxicology of Chemicals. The Role of Bioaccumulation in Environmental Risk Assessment: The Aquatic Environment and Related Food Webs. (1995). Technical report 67, Brussel, Belgium.

115. Organization for Economic Cooperation and Development, Bioconcentration: Flow-through fish test. OECD Guide-Line for Testing of Chemicals: Draft Guideline 305. Paris, France, 1994.

116. Nendza M. Structure-Activity Relationships in Environmental Sciences, Chapman & Hall, London, 1998.

117. Pavan M, Worth AP, Netzeva TI. Review of QSAR Models for Bioconcentration. (2006). EUR 22327 EN.

118. Oliver JA. Opportunities for using fewer animals in acute toxicity studies. In: Proceedings from the International Seminar, Chemicals Testing and Animal Welfare. The National Chemicals Inspectorate. Stockholm, Sweden, May 20-22, 1986;1:19-142

119. Ames BN, Lee FD, Durston WE. An improved bacterial test system for the detection and classification of mutagens and carcinogens. Proc Natl Acad Sci U S A. 1973 Mar;70(3):782-6

120. Ames BN, Durston WE, Yamasaki E, Lee FD. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. Proc Natl Acad Sci U S A. 1973 Aug;70(8):2281-5

121. Fu W, Franco A, Trapp S. Methods for estimating the bioconcentration factor of ionizable organic chemicals. Environ Toxicol Chem. 2009 Jul;28(7):1372-9.

122. EC Project FOOTPRINT: creating tools for pesticide risk assessment and management in Europe - http://sitem.herts.ac.uk/aeru/footprint/

123. Cefic LRI BCF database - http://www.cefic-lri.org/lri-toolbox/bcf

124. Arnot JA, Gobas FAPC, A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ Rev. 2006 Dec 1;14(4):257-297.

125. U.S. Enviromental Protection Agency - Toxicity Estimation Software Tool (T.E.S.T.) - http://www.epa.gov/ordntrnt/ORD/NRMRL/std/qsar/qsar.html

126. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N, Müller KR. Benchmark data set for in silico prediction of Ames mutagenicity. J Chem Inf Model. 2009 Sep;49(9):2077-81.

127. U.S. Nationl Library of Medicine. Chemical Carcinogenesis Research Information System (CCRIS) TOXNET database. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS

128. Helma C, Cramer T, Kramer S, De Raedt L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. J Chem Inf Comput Sci. 2004 Jul-Aug;44(4):1402-11.

129. Feng J, Lurati L, Ouyang H, Robinson T, Wang Y, Yuan S, Young SS. Predictive toxicology: benchmarking molecular descriptors and statistical methods. J Chem Inf Comput Sci. 2003 Sep-Oct;43(5):1463-70

130. Judson PN, Cooke PA, Doerrer NG, Greene N, Hanzlik RP, Hardy C, Hartmann A, Hinchliffe D, Holder J, Müller L, Steger-Hartmann T, Rothfuss A, Smith M, Thomas K, Vessey JD, Zeiger E. Towards the creation of an international toxicology information centre. Toxicology. 2005 Sep 15;213(1-2):117-28.

131. U.S. National Library of Medicine. Genetic Toxicology Data Bank (GENE-TOX) TOXNET database. http://toxnet.nlm.nih.gov/newtoxnet/genetox.htm

132. ACD/Labs ToxSuite. Advanced Chemistry Development Inc., Toronto, Canada; software available at http://www.acdlabs.com/

133. TerraQSAR. TerraBase Inc., Hamilton, Canada; software available at http://www.terrabase-inc.com/

134. T. Martin, Toxicity Estimation Software Tool (TEST); software available at http://www.epa.gov/nrmrl/std/qsar/qsar.html

135. TOxicity Prediction by Komputer Assisted Technology (TOPKAT). Accelrys Inc., San Diego, USA; software available at http://accelrys.com/

136. Lombardo A, Roncaglioni A, Boriani E, Milan C, Benfenati E. Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. Chem Cent J. 2010 Jul 29;4 Suppl 1:S1.

137. Ferrari T, Gini G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. Chem Cent J. 2010 Jul 29;4 Suppl 1:S2.

138. ChemOffice. PerkinElmer Inc., Massachusetts, USA; software available at https://www.cambridgesoft.com/

139. Instant JChem. ChemAxon Kft., Budapest, Hungary; software available at http://www.chemaxon.com/

140. In-Silico Tools. Kode s.r.l., Pisa, Italy; http://kode-solutions.net/

141. The Chemistry Development Kit - http://sourceforge.net/projects/cdk/

142. SMiles ARbitrary Target Specification (SMARTS) - http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

143. Open Babel: The Open Source Chemistry Toolbox - http://openbabel.org/

144. Open Babel: Canonical Coding Algorithm - http://openbabel.org/dev-api/canonical_code_algorithm.shtml

145. Accelrys Software Inc., Discovery Studio Modeling Environment, Release 3.0, San Diego: Accelrys Software Inc.

Thesis annexes

Annex A.
SARpy fragments

I. List of fragments extracted by SARpy

In this annex are reported the complete lists of fragments extracted by SARpy. The fragments have been sorted primarily by the target property, then by their likelihood ratio (LR), in decreasing order. The "inf" LR value means that the fragment correctly predicted all the molecules targeted in the training set. For example, a fragment associated with true positive (TP) molecules, which was found only in TP molecules.

Table A. Fragments extracted by SARpy and related to the true positive (TP) and false positive (FP) predictions obtained by the CAESAR model for mutagenicity. These fragments were obtained using a training set composed only by TP and FP molecules. The SMARTS column contain the SMILES representation of the fragments, Target is the property statistically associated to the fragment and Training LR is the likelihood ratio calculated by SARpy.

SMARTS	Target	Training LR
c1ccc(o1)[N](=O)O	TP	inf
N#N	TP	inf
c1csc(c1)	TP	inf
O=NN(C(=O))C	TP	inf
N1C(C1c1ccccc1)	TP	inf
CCC(Nc1ccnc2c1ccc(c2))	TP	inf
COS(=O)(=O)	TP	inf
C(COC(C))CN	TP	inf
C1=Cc2c3c1cccc3ccc2	TP	inf
c1ccc2c(c1C)ncc(n2)	TP	inf
Cc1cn(c2c1cccc2)	TP	inf
c1snc2c1cc(cc2)[N](=O)O	TP	inf
O=C(c1ccc(cc1)[N](=O)O)	TP	inf
CCCN(CCC)N=O	TP	inf
C(Cn1cncc1)	TP	inf
[N](=O)c1cc2ccccc2c2c1cccc2	TP	inf

SMARTS	Target	Training LR
C=Cc1ccc(cc1)[N]O	TP	inf
N(c1ccc(cc1)O)C	TP	inf
c1cccc2c1ccc1c2cccn1	TP	inf
c1cc2ccccc2c2c1cc(cc2)[N]	TP	inf
ClC=C(S)	TP	inf
O[N]c1ccc2c(c1)cc[nH]2	TP	inf
COc1ccc(cc1)N=O	TP	inf
COCCBr	TP	inf
C(c1ccc2c3c1ccc1c3c(cc2)ccc1)O	TP	inf
Nc1ccc(c(c1)N)N=Nc1ccccc1	TP	inf
c1cccc2c1oc1ccccc1c2=O	TP	inf
Cc1ccc2c(c1)cc1c(n2)cccc1	TP	inf
Sc1ccc(cc1)[N]O	TP	inf
ClC(C(=O))Cl	TP	inf
CN([N](=O)O)	TP	inf
Oc1ccc(c2c1C(=O)c1ccccc1C2=O)O	TP	inf
[N]c1cc(ccc1c1ccccc1)[N]	TP	inf
Nc1ccc2c(c1)nccc2	TP	inf
[N](c2cc(cc(c2C)[N])[N])	TP	inf
Nc1ncnc(c1)	TP	inf
C(=O)OCC1CO1	TP	inf
CCOCNN	TP	inf
COCC(=CCl)	TP	inf
N(CNC)CC	TP	inf
OC=CC=C	TP	inf
c1cc(nn1)	TP	inf
c1ccc(cc1)c1ccc(cc1)OC	TP	inf
c1ccc2c(c1)c1ccccc1C2C	TP	inf
Clc1ccc(c(c1))[N](=O)O	TP	inf
c1ccc(nn1)NN	TP	inf
BrC(=C)	TP	inf
CC(ON(C(=O)c1ccccc1)OC(=O)C)	TP	inf
c1nccc2c1[nH]c1c2cccc1	TP	inf
Nc1scc(n1)c1ccc(cc1)	TP	inf
C(=O)c1ccccc1Cl	TP	inf
O=Nc1ccc(cc1)c1ccccc1	TP	13.23
O(Cc1cccc2c1cccc2)C	TP	9.45
[N]c1ccc(cc1)c1ccc(cc1)[N]	TP	8.64
Cc1ccc2c(c1)cc1c(c2)cccc1	TP	7.29
C=Cc1ccc(cc1)[N]	TP	7.29
[N]c1ccc(cc1)Oc1ccccc1	TP	6.21
c1ccc2c(c1)cc1c(n2)cccc1	TP	6.21
CCN(C)N=O	TP	5.47
O=Cc1cccc(c1)[N]O	TP	5.4
O=Nc1cccc2c1cccc2	TP	4.66

SMARTS	Target	Training LR
c1cccc2c1snc2	TP	4.59
ClCCNCC	TP	4.32
CCN(N=O)CC	TP	4.14
[N]c1ccc2c(c1)c1ccccc1cc2	TP	4.14
NCc1ccc(c(c1)O)	TP	4.05
c1cc(O)cc2c1nccc2	TP	4.05
Cc1ccc(c(c1)[N])C	TP	3.78
n1cnc(c1c1ccccc1)	TP	3.78
c1nc2cc(cc(c2nc1))[N]	TP	3.78
C(c1ccccc1)OCc1cccc(c1)	TP	3.51
O[N]c1ccc2c(c1)cccc2	TP	3.46
C1C=Cc2c(C1)ccc1c2cccc1	TP	3.24
c1cc(O)c(cc1O)	TP	3.24
[N](=O)c1cc(ccc1)c1ccccc1	TP	3.2
[N]c1cc(ccc1)c1ccccc1	TP	3.14
BrCBr	TP	2.97
O=CCCCN	TP	2.84
[N]c1ccc(cc1)c1ccccc1	TP	2.76
C(c1cc2ccc3c4c2c(c1)ccc4ccc3)O	TP	2.7
Fc1ccc(cc1)	TP	2.5
c1ccc(cc1)N=Nc1ccc(cc1)[N](=O)O	TP	2.43
CC(=CC1C(C1))	TP	2.43
COP(Oc1ccc(cc1))	TP	2.16
[N]c1ccc(c(c1)[N](=O)O)N	TP	2.16
O=Cc1cccc(c1)[N]	TP	2.16
[N]c1nc2c(n1C)cccc2	TP	2.16
OOC	TP	2.16
n1c2ccccc2c2c1cccc2	TP	2.09
c1ccc2c3c1ccc1c3c(cc2)ccc1	TP	2.04
ClC(Cl)Cl	TP	2.03
c1ccc(cc1)OCC1OC1	TP	2.03
c1c2ccccc2cc2c1cccc2	TP	1.82
ClCC=C	TP	1.8
O[N](=O)C	TP	1.62
CN(C)C	TP	1.59
CN(C(C)C)C	TP	1.57
c1cc2cccc3c2c(c1)c1ccccc31	TP	1.49
CCNCC[N]	TP	1.47
c1ccc(cc1C(C))c1ccccc1	TP	1.41
[N](c1ccc(c(c1))C)	TP	1.4
c1ccc(cc1)N=Nc1ccc(cc1)[N]	TP	1.38
c1ccc2c3c1ccc1c3c(cc2)c(cc1)O	TP	1.35
CCCCc1ccc(cc1)	TP	1.33
N(c1c(cc(cc1)))[N]	TP	1.28
c1nsc(c1)	TP	1.28

SMARTS	Target	Training LR
OCCc1ccc(cc1)	TP	1.27
c1ccc(cc1)C(=O)c1cc(O)c(c(c1))	TP	1.24
C(O)O	TP	1.24
C[N](C)	TP	1.24
c1cc2ccccc2c2c1c1ccccc1cc2	TP	1.22
NCCC	TP	1.2
Nc1ccc(cc1)Cc1ccc(cc1)	TP	1.15
[N](=O)c1ccc(cc1)[N]	TP	1.14
Nc2ccccn2	TP	1.08
COc1ccc(cc1)C(C)C	TP	1.08
O[N](=O)c1cccc(c1)[N]	TP	1.08
c1nc2c(s1)cccc2	TP	1.08
BrCCC	TP	1.08
[N]c1cccc(c1)[N]	TP	1.07
C(=O)c1ccccc1C	TP	1.07
c1c(O)ccc2c1cccc2	TP	1.06
O(Cc1ccccc1)C	TP	1.04
OCCCl	TP	1.01
COC(=O)C	TP	1.01
n1ncc(c(c1=O)Cl)Cl	FP	inf
CCCCCCCC1CO1	FP	24.07
O=C1C=CCC=C1	FP	13.58
CCCCCC1CO1	FP	13.58
Oc1ccccc1NC(=O)C	FP	11.11
C(F)(F)F	FP	9.87
CNNC	FP	9.87
CCCCCl	FP	8.02
N(CSCC)	FP	7.41
N(c1ccccc1)c1ccccc1	FP	6.48
N=[N]O	FP	5.55
CCCCCCCC	FP	4.23
[N](=O)c1ccccc1O	FP	4.14
CCCCCC=C	FP	3.09
CCc1ccc(c(c1)OC)	FP	2.96
[N](c1cccc(c1C)[N]O)	FP	2.88
NCC(C(C(CO))O)	FP	2.88
CCOCCOCC	FP	2.69
CCN(c1ccc(cc1)N=N)CC	FP	2.59
c1ccc(cc1)Br	FP	2.56
[N]c1ccccc1O	FP	2.36
OCc1cccc(c1)[N]	FP	2.28
Clc1ccc(c(c1)[N])	FP	2.18
Oc1ccc(c(c1)O)C(=O)	FP	2.18
CCc1cccc(c1)[N]	FP	2.08
C(=O)COCC	FP	1.85

SMARTS	Target	Training LR
C(C(=O)O)Cl	FP	1.85
c1ccc2c(c1)oc(=O)cc2	FP	1.85
c1ccc(cc1)C1OC1C	FP	1.79
O=C(c1ccccc1)O	FP	1.67
C(=O)Nc1ccccc1	FP	1.48
CC(=O)N	FP	1.47
Cc1cccnc1	FP	1.44
CCOCCN	FP	1.4
c1ccc(c(c1)O)C	FP	1.34
C(=O)C[N]	FP	1.26
OCc1ccccc1	FP	1.15
S(=O)(=O)O	FP	1.09
c1ccc(c(c1))C(=O)c1ccccc1	FP	1.09
N(Cc1ccccc1)C	FP	1.04

Table B. Fragments extracted by SARpy and related to the true negative (TN) and false negative (FN) predictions obtained by the CAESAR model for mutagenicity. These fragments were obtained using a training set composed only of TN and FN molecules. The SMARTS column contain the SMILES representation of the fragments, Target is the property statistically associated to the fragment and Training LR is the likelihood ratio calculated by SARpy.

SMARTS	Target	Training LR
CCCC=CCCCC	TN	inf
C(=O)NC(C)C	TN	inf
CCCOC(c1ccccc1)	TN	inf
CCCC(CCO)O	TN	inf
C(=O)COc1ccccc1	TN	inf
Oc1ccc(cc1)C=C	TN	inf
CN=C	TN	inf
c1cc(ccc1S(=O)(=O))[N]	TN	inf
c1cc(Cl)c(c(c1)Cl)Cl	TN	inf
c1ccc(cc1)C(c1ccccc1)(C)	TN	inf
NC(=O)Nc1ccccc1	TN	inf
Brc1cc(Br)c(c(c1))	TN	inf
O=CCCC(=O)O	TN	inf
C(=C)CCC=C(C)	TN	inf
COP(=S)(O)	TN	inf
Nc1ccccn1	TN	inf
n1sccc1	TN	inf
Cc1cc(c(c(c1)C(C)(C)C))	TN	inf
c1ccc2c(c1)cc(n2)C	TN	inf
CCCC(=O)NC	TN	inf
C(c1ccccc1)C(N)C	TN	inf
FCC(F)(F)	TN	inf
CCCCS	TN	inf
S(=O)(=O)c1cccc2c1nccc2	TN	inf
CC(COC(=O)CC)C	TN	inf
CN1CCNCC1	TN	inf
OCC1OC(CC1O)n1ccc(nc1=O)	TN	inf
CCCCCCCCOC(=O)C	TN	inf
c1coc2c(c1)c(O)cc(c2)	TN	inf
OCCOC(=O)C=C	TN	inf
CSc1ccc(cc1)	TN	inf
C1C2CCC(C1)C2	TN	inf
c1ncnc(n1)	TN	inf
c1ccc(cc1)CCCCCCCC	TN	inf
CCCCCCCCCCCN	TN	inf
CCCCOP(O)O	TN	inf
Cc1ccncc1	TN	inf

SMARTS	Target	Training LR
c1scnc1	TN	inf
CC(=C)C#N	TN	inf
c1cccc2c1nc(o2)c1ccccc1	TN	inf
c1cnccn1	TN	inf
C[Si]	TN	inf
O=C(c1cccc2c1nccc2)O	TN	inf
c1nncc2c1cncc2	TN	inf
c1cnn(c(=O)c1)	TN	inf
OC(=O)CC(=O)	TN	inf
c1ccc(cc1)OP(Oc1ccccc1)O	TN	inf
c1ccc(cc1)C(=O)c1ccc(cc1)	TN	inf
COC(=O)c1ccc(cc1)C	TN	inf
CNS(=O)(=O)c1ccc(cc1)	TN	inf
Oc1ccc2c(c1)cc(n2)	TN	inf
CCC=CC=CC=O	TN	inf
CC(COS(=O)(=O))	TN	inf
c1cccs1	TN	inf
C(=O)OC(=O)	TN	inf
c1ccc(cc1Cl)Cl	TN	10.18
CCSC	TN	5.77
COC(C)(C)C	TN	4.1
Oc1ccc(c(c1)Cl)	TN	3.95
OCCOCCOC	TN	3.8
OCCN(CC)C	TN	3.11
C(=O)OCCC	TN	2.99
c1ccc(cc1C)Cl	TN	2.89
Oc1ccc(cc1)C(=O)	TN	2.35
Clc1ccc(cc1)	TN	2.29
CCCCOc1ccc(cc1)	TN	2.28
OC(=O)CO	TN	2.18
CC(CCC=C(C))C	TN	2.05
COc1cccc(c1)C	TN	1.63
NCCCN	TN	1.59
COc1cccc(c1)C(=O)	TN	1.52
C(=O)c1cccc(c1[N])	TN	1.44
Cc1ccc(cc1)Cl	TN	1.4
C(=O)Nc1ccccc1	TN	1.39
OCc1ccc(cc1)Cl	TN	1.22
N#Cc1ccc(cc1)	TN	1.22
CNCCCO	TN	1.09
N#C	TN	1.04
ONC(=O)c1ccccc1	FN	inf
c1nc(c([nH]1)c1ccccc1)c1ccccc1	FN	inf
CC12OOC1c1c(O2)cccc1	FN	52.67
COO	FN	16.46

SMARTS	Target	Training LR
CC(=O)NO	FN	9.88
[N](=O)	FN	3.66
CC(=CC)c1ccc(cc1)	FN	3.29
ClCCl	FN	2.71
CCCl	FN	2.44
COCCOC(=O)	FN	1.98
NCN	FN	1.86
Cc1ccc(cc1)[N]	FN	1.76
CC(CCC(O)(C)C)	FN	1.54
CC1CCC(O1)	FN	1.32
c1ccc(cc1)OC	FN	1.23
C(Cc1ccccc1)C	FN	1.21
c1ccc(cc1)[N]	FN	1.21
CCOCC(CO)	FN	1.1

II. Statistical analysis of the fragments

The ruleset generated by SARpy were used to predict the molecules of both training and prediction sets. The occurrences of each fragment were calculated, as well as the number of correct assignment (e.g. a fragment related to true positive predictions, found in molecule whose CAESAR prediction was a TP), wrong assignment, and the percentage of correctness. The results are reported in the following tables, grouped by the fragment's target property. The fragments are sorted by their LR values in decreasing order.

Table C. Total occurrences and percentage of correct assignments, for the TRUE POSITIVE fragments extracted by SARpy from the TP-FP dataset. The results are shown for both training and prediction set, as evaluated using either the complete SARpy ruleset or the "Single Fragments" approach (each fragment considered separately).

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
c1ccc(o1)[N](=O)O	inf	55	100%	55	100%	23	100%	23	100%
N#N	inf	40	100%	40	100%	25	92%	25	92%
c1csc(c1)	inf	32	100%	32	100%	13	85%	13	85%
O=NN(C(=O))C	inf	31	100%	31	100%	15	93%	15	93%
N1C(C1c1ccccc1)	inf	29	100%	29	100%	7	86%	7	86%
CCC(Nc1ccnc2c1ccc(c2))	inf	23	100%	23	100%	8	100%	8	100%
COS(=O)(=O)	inf	23	100%	23	100%	11	82%	11	82%
C(COC(C))CN	inf	21	100%	23	100%	12	67%	12	67%
C1=Cc2c3c1cccc3ccc2	inf	20	100%	20	100%	5	100%	5	100%
c1ccc2c(c1C)ncc(n2)	inf	19	100%	19	100%	4	100%	4	100%
Cc1cn(c2c1cccc2)	inf	18	100%	18	100%	6	50%	6	50%
C(Cn1cncc1)	inf	17	100%	19	100%	11	91%	14	93%
c1snc2c1cc(cc2)[N](=O)O	inf	17	100%	17	100%	6	100%	6	100%
CCCN(CCC)N=O	inf	17	100%	18	100%	14	93%	14	93%
O=C(c1ccc(cc1)[N](=O)O)	inf	17	100%	18	100%	5	100%	6	100%
[N](=O)c1cc2ccccc2c2c1cccc2	inf	16	100%	16	100%	17	88%	17	88%
C=Cc1ccc(cc1)[N]O	inf	16	100%	17	100%	6	100%	7	100%
N(c1ccc(cc1)O)C	inf	16	100%	18	100%	8	88%	10	90%
c1cccc2c1ccc1c2cccn1	inf	15	100%	15	100%	10	80%	10	80%
c1cc2ccccc2c2c1cc(cc2)[N]	inf	13	100%	16	100%	6	100%	8	100%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
ClC=C(S)	inf	13	100%	13	100%	3	100%	3	100%
COc1ccc(cc1)N=O	inf	12	100%	13	100%	4	75%	4	75%
COCCBr	inf	12	100%	12	100%	5	100%	5	100%
O[N]c1ccc2c(c1)cc[nH]2	inf	12	100%	12	100%	3	100%	4	100%
C(c1ccc2c3c1ccc1c3c(cc2)ccc1)O	inf	11	100%	11	100%	5	80%	6	83%
c1cccc2c1oc1ccccc1c2=O	inf	10	100%	10	100%	7	100%	7	100%
Cc1ccc2c(c1)cc1c(n2)cccc1	inf	10	100%	14	100%	8	75%	10	80%
ClC(C(=O))Cl	inf	10	100%	10	100%	2	50%	2	50%
CN([N](=O)O)	inf	10	100%	10	100%	2	100%	2	100%
Nc1ccc(c(c1)N)N=Nc1ccccc1	inf	10	100%	11	100%	2	100%	4	100%
Sc1ccc(cc1)[N]O	inf	10	100%	10	100%	1	100%	1	100%
[N](c2cc(cc(c2C)[N])[N])	inf	9	100%	9	100%	2	100%	2	100%
[N]c1cc(ccc1c1ccccc1)[N]	inf	9	100%	10	100%	6	100%	6	100%
C(=O)OCC1CO1	inf	9	100%	10	100%	4	100%	4	100%
CCOCNN	inf	9	100%	15	100%	4	100%	9	100%
Nc1ccc2c(c1)nccc2	inf	9	100%	16	100%	6	83%	9	78%
Nc1ncnc(c1)	inf	9	100%	21	100%	9	67%	14	71%
Oc1ccc(c2c1C(=O)c1ccccc1C2=O)O	inf	9	100%	19	100%	7	100%	12	92%
c1cc(nn1)	inf	8	100%	9	100%	5	100%	6	100%
COCC(=CCl)	inf	8	100%	8	100%	8	88%	8	88%
N(CNC)CC	inf	8	100%	19	100%	4	100%	9	100%
OC=CC=C	inf	8	100%	9	100%	6	100%	6	100%
c1ccc(cc1)c1ccc(cc1)OC	inf	7	100%	7	100%	2	50%	2	50%
c1ccc2c(c1)c1ccccc1C2C	inf	7	100%	7	100%	5	80%	5	80%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
Clc1ccc(c(c1))[N](=O)O	inf	7	100%	8	100%	5	60%	5	60%
BrC(=C)	inf	6	100%	9	100%	3	100%	6	100%
c1ccc(nn1)NN	inf	6	100%	7	100%	2	100%	2	100%
CC(ON(C(=O)c1ccccc1)OC(=O)C)	inf	4	100%	4	100%	3	100%	4	100%
C(=O)c1ccccc1Cl	inf	3	100%	4	100%	4	100%	5	80%
c1nccc2c1[nH]c1c2cccc1	inf	3	100%	4	100%	2	100%	3	100%
Nc1scc(n1)c1ccc(cc1)	inf	3	100%	3	100%	1	100%	1	100%
O=Nc1ccc(cc1)c1ccccc1	13.23	28	96%	50	98%	18	94%	29	97%
O(Cc1cccc2c1cccc2)C	9.45	27	96%	36	97%	19	95%	20	95%
[N]c1ccc(cc1)c1ccc(cc1)[N]	8.64	20	95%	33	97%	7	100%	14	100%
Cc1ccc2c(c1)cc1c(c2)cccc1	7.29	19	95%	28	96%	16	81%	19	84%
C=Cc1ccc(cc1)[N]	7.29	11	91%	28	96%	5	100%	13	100%
[N]c1ccc(cc1)Oc1ccccc1	6.21	14	93%	24	96%	7	100%	11	91%
c1ccc2c(c1)cc1c(n2)cccc1	6.21	25	92%	72	96%	12	83%	39	85%
CCN(C)N=O	5.47	28	100%	85	95%	17	94%	48	94%
O=Cc1cccc(c1)[N]O	5.4	13	92%	21	95%	12	100%	16	94%
O=Nc1cccc2c1cccc2	4.66	45	91%	73	95%	30	87%	64	91%
c1cccc2c1snc2	4.59	18	94%	18	94%	7	86%	7	86%
ClCCNCC	4.32	15	87%	34	94%	7	100%	17	94%
[N]c1ccc2c(c1)c1ccccc1cc2	4.14	12	100%	49	94%	5	100%	36	97%
CCN(N=O)CC	4.14			49	94%			30	93%
NCc1ccc(c(c1)O)	4.05	10	100%	16	94%	3	100%	3	100%
c1cc(O)cc2c1nccc2	4.05	9	89%	16	94%	2	100%	3	100%
n1cnc(c1c1ccccc1)	3.78	10	90%	15	93%	6	100%	6	100%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
c1nc2cc(cc(c2nc1))[N]	3.78	8	88%	15	93%	8	100%	10	100%
Cc1ccc(c(c1)[N])C	3.78	7	86%	15	93%	4	100%	9	89%
C(c1ccccc1)OCc1cccc(c1)	3.51	10	90%	14	93%	5	60%	9	78%
O[N]c1ccc2c(c1)cccc2	3.46	13	85%	69	93%	7	86%	56	93%
C1C=Cc2c(C1)ccc1c2cccc1	3.24	14	93%	26	92%	3	67%	16	75%
c1cc(O)c(cc1O)	3.24	14	86%	39	92%	1	100%	16	94%
[N](=O)c1cc(ccc1)c1ccccc1	3.2	10	100%	77	92%	2	100%	53	92%
[N]c1cc(ccc1)c1ccccc1	3.14	11	82%	101	92%	2	100%	60	93%
BrCBr	2.97	8	88%	12	92%	5	40%	5	40%
O=CCCCN	2.84	9	89%	23	91%	6	50%	19	74%
[N]c1ccc(cc1)c1ccccc1	2.76	16	94%	101	91%	21	86%	63	92%
C(c1cc2ccc3c4c2c(c1)ccc4ccc3)O	2.7	2	100%	11	91%	1	100%	3	67%
Fc1ccc(cc1)	2.5	23	83%	41	90%	16	69%	25	76%
CC(=CC1C(C1))	2.43	1	100%	10	90%	1	100%	7	100%
c1ccc(cc1)N=Nc1ccc(cc1)[N](=O)O	2.43	1	0%	10	90%	3	100%	7	100%
COP(Oc1ccc(cc1))	2.16	8	100%	9	89%	2	50%	2	50%
O=Cc1cccc(c1)[N]	2.16	11	91%	45	89%	3	100%	27	89%
OOC	2.16	9	89%	9	89%	6	50%	9	67%
[N]c1ccc(c(c1)[N](=O)O)N	2.16	8	88%	18	89%	4	75%	11	91%
[N]c1nc2c(n1C)cccc2	2.16	21	86%	36	89%	12	83%	18	89%
n1c2ccccc2c2c1cccc2	2.09	13	85%	35	89%	3	33%	14	86%
c1ccc2c3c1ccc1c3c(cc2)ccc1	2.04	45	76%	120	88%	16	75%	50	88%
c1ccc(cc1)OCC1OC1	2.03	12	83%	17	88%	3	67%	7	86%
ClC(Cl)Cl	2.03	12	83%	17	88%	6	33%	7	43%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
c1c2ccccc2cc2c1cccc2	1.82	45	84%	147	87%	24	92%	73	88%
ClCC=C	1.8	11	91%	23	87%	2	100%	15	93%
O[N](=O)C	1.62	7	71%	14	86%	7	71%	14	86%
CN(C)C	1.59	28	71%	124	85%	18	72%	71	76%
CN(C(C)C)C	1.57			34	85%			24	75%
c1cc2cccc3c2c(c1)c1ccccc31	1.49	19	74%	39	85%	7	100%	26	96%
CCNCC[N]	1.47	9	89%	45	84%			23	78%
c1ccc(cc1C(C))c1ccccc1	1.41	21	90%	87	84%	8	75%	43	79%
[N](c1ccc(c(c1))C)	1.4	56	88%	173	84%	20	80%	71	83%
c1ccc(cc1)N=Nc1ccc(cc1)[N]	1.38	18	83%	49	84%	7	71%	20	85%
c1ccc2c3c1ccc1c3c(cc2)c(cc1)O	1.35			12	83%			7	71%
CCCCc1ccc(cc1)	1.33	4	100%	95	83%	4	75%	48	71%
c1nsc(c1)	1.28	11	100%	69	83%	3	67%	22	73%
N(c1c(cc(cc1)))[N]	1.28	7	71%	23	83%	2	100%	4	100%
OCCc1ccc(cc1)	1.27	20	70%	183	83%	7	100%	114	75%
C(O)O	1.24	10	80%	84	82%	2	100%	52	81%
C[N](C)	1.24	53	77%	482	82%	16	81%	224	84%
c1ccc(cc1)C(=O)c1cc(O)c(c(c1))	1.24			56	82%	1	0%	29	90%
c1cc2ccccc2c2c1c1ccccc1cc2	1.22	19	89%	77	82%	5	100%	27	93%
NCCC	1.2	2	100%	273	82%	3	33%	146	82%
Nc1ccc(cc1)Cc1ccc(cc1)	1.15			21	81%			8	88%
[N](=O)c1ccc(cc1)[N]	1.14	5	100%	47	81%	4	100%	32	94%
O[N](=O)c1cccc(c1)[N]	1.08	5	100%	90	80%	5	60%	47	85%
Nc2ccccn2	1.08	4	100%	10	80%	2	100%	6	83%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
c1nc2c(s1)cccc2	1.08	11	82%	15	80%	3	33%	5	40%
BrCCC	1.08	14	79%	35	80%	14	71%	18	72%
COc1ccc(cc1)C(C)C	1.08	1	0%	10	80%			9	100%
[N]c1cccc(c1)[N]	1.07	3	100%	119	80%	3	100%	57	86%
C(=O)c1ccccc1C	1.07	4	100%	114	80%	3	100%	67	81%
c1c(O)ccc2c1cccc2	1.06	6	83%	74	80%	1	100%	35	77%
O(Cc1ccccc1)C	1.04			131	79%			63	70%
OCCCl	1.01	9	100%	38	79%	9	56%	25	68%
COC(=O)C	1.01	7	100%	142	79%	5	60%	70	81%

Table D. Total occurrences and percentage of correct assignments, for the FALSE POSITIVE fragments extracted by SARpy from the TP-FP dataset. The results are shown for both training and prediction set, as evaluated using either the complete SARpy ruleset or the "Single Fragments" approach (each fragment considered separately).

		Training Set				**Prediction Set**			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
n1ncc(c(c1=O)Cl)Cl	inf	7	100%	7	100%	1	0%	1	0%
CCCCCCCC1CO1	24.07	14	93%	15	87%	8	50%	8	50%
O=C1C=CCC=C1	13.58	13	85%	14	79%	6	33%	6	33%
CCCCCC1CO1	13.58	12	75%	28	79%	6	83%	15	67%
Oc1ccccc1NC(=O)C	11.11	8	75%	8	75%	2	50%	2	50%
C(F)(F)F	9.87	9	89%	11	73%	6	50%	8	38%
CNNC	9.87	10	80%	11	73%	3	67%	4	50%
CCCCCl	8.02	14	71%	19	68%	5	60%	5	60%
N(CSCC)	7.41	15	67%	15	67%	10	30%	10	30%
N(c1ccccc1)c1ccccc1	6.48	11	64%	11	64%	12	25%	12	25%
N=[N]O	5.55	12	75%	15	60%	5	40%	5	40%
CCCCCCCC	4.23	34	38%	75	53%	20	0%	41	24%
[N](=O)c1ccccc1O	4.14	24	67%	36	53%	9	44%	12	42%
CCCCCC=C	3.09	4	50%	33	45%	4	50%	26	23%
CCc1ccc(c(c1)OC)	2.96	13	62%	27	44%	3	0%	11	18%
[N](c1cccc(c1C)[N]O)	2.88	5	80%	16	44%	3	0%	8	13%
NCC(C(C(CO))O)	2.88	4	75%	16	44%			6	17%
CCOCCOCC	2.69	6	83%	19	42%	3	33%	9	22%
CCN(c1ccc(cc1)N=N)CC	2.59	15	47%	17	41%	1	100%	3	33%
c1ccc(cc1)Br	2.56	13	62%	22	41%	7	14%	11	18%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
[N]c1ccccc1O	2.36	32	38%	95	39%	9	33%	36	25%
OCc1cccc(c1)[N]	2.28	6	100%	21	38%	2	100%	10	30%
Clc1ccc(c(c1)[N])	2.18	13	54%	27	37%	9	44%	19	37%
Oc1ccc(c(c1)O)C(=O)	2.18	19	47%	27	37%	8	13%	10	20%
CCc1cccc(c1)[N]	2.08	4	75%	25	36%			9	44%
C(=O)COCC	1.85	5	80%	24	33%	1	0%	16	38%
C(C(=O)O)Cl	1.85	11	55%	21	33%	7	43%	14	21%
c1ccc2c(c1)oc(=O)cc2	1.85	15	40%	24	33%	10	30%	13	31%
c1ccc(cc1)C1OC1C	1.79	17	53%	43	33%	7	43%	25	44%
O=C(c1ccccc1)O	1.67	13	62%	45	31%	7	29%	25	40%
C(=O)Nc1ccccc1	1.48	31	55%	112	29%	10	20%	55	15%
CC(=O)N	1.47	40	45%	201	28%	21	48%	96	23%
Cc1cccnc1	1.44	13	38%	25	28%	9	33%	18	22%
CCOCCN	1.4	4	50%	40	28%			18	28%
c1ccc(c(c1)O)C	1.34	22	32%	128	27%	14	36%	65	22%
C(=O)C[N]	1.26	6	50%	71	25%	6	0%	41	22%
OCc1ccccc1	1.15	8	50%	278	24%	4	75%	153	29%
c1ccc(c(c1))C(=O)c1ccccc1	1.09	8	50%	88	23%	10	60%	55	18%
S(=O)(=O)O	1.09	2	0%	44	23%	1	100%	16	31%
N(Cc1ccccc1)C	1.04			91	22%			34	21%

Table E. Total occurrences and percentage of correct assignments, for the TRUE NEGATIVE fragments extracted by SARpy from the TN-FN dataset. The results are shown for both training and prediction set, as evaluated using either the complete SARpy ruleset or the "Single Fragments" approach (each fragment considered separately).

		Training Set				**Prediction Set**			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
Brc1cc(Br)c(c(c1))	inf	13	100%	14	100%	4	100%	5	100%
C(=C)CCC=C(C)	inf	13	100%	21	100%	4	75%	6	83%
C(=O)COc1ccccc1	inf	23	100%	30	100%	5	100%	9	100%
C(=O)NC(C)C	inf	45	100%	45	100%	28	89%	29	90%
C(=O)OC(=O)	inf	4	100%	9	100%			1	100%
C(c1ccccc1)C(N)C	inf	10	100%	23	100%	3	100%	6	100%
C[Si]	inf	5	100%	5	100%	3	67%	3	67%
C1C2CCC(C1)C2	inf	8	100%	8	100%	1	100%	1	100%
c1cc(ccc1S(=O)(=O))[N]	inf	17	100%	17	100%	14	100%	14	100%
c1cc(Cl)c(c(c1)Cl)Cl	inf	17	100%	17	100%	12	100%	12	100%
c1ccc(cc1)C(=O)c1ccc(cc1)	inf	4	100%	6	100%	2	50%	2	50%
c1ccc(cc1)C(c1ccccc1)(C)	inf	15	100%	18	100%	5	100%	5	100%
c1ccc(cc1)CCCCCCCC	inf	7	100%	12	100%	6	100%	8	100%
c1ccc(cc1)OP(Oc1ccccc1)O	inf	3	100%	4	100%	1	100%	3	100%
c1ccc2c(c1)cc(n2)C	inf	11	100%	17	100%	4	75%	4	75%
c1cccc2c1nc(o2)c1ccccc1	inf	6	100%	6	100%	3	100%	3	100%
c1cccs1	inf	3	100%	7	100%	1	100%	2	100%
c1cnccn1	inf	6	100%	10	100%	5	80%	5	80%
c1cnn(c(=O)c1)	inf	5	100%	5	100%	1	100%	1	100%
c1coc2c(c1)c(O)cc(c2)	inf	8	100%	9	100%	6	83%	6	83%
c1ncnc(n1)	inf	8	100%	8	100%	4	75%	4	75%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
c1nncc2c1cncc2	inf	4	100%	4	100%	1	100%	1	100%
c1scnc1	inf	7	100%	8	100%	5	40%	9	56%
CC(=C)C#N	inf	7	100%	9	100%	2	0%	3	33%
CC(COC(=O)CC)C	inf	9	100%	18	100%	5	100%	11	100%
CC(COS(=O)(=O))	inf	3	100%	3	100%				
Cc1cc(c(c(c1)C(C)(C)C))	inf	11	100%	14	100%	6	100%	7	100%
Cc1ccncc1	inf	7	100%	7	100%	1	100%	1	100%
CCC=CC=CC=O	inf	3	100%	10	100%	2	100%	3	100%
CCCC(=O)NC	inf	11	100%	28	100%	7	86%	18	83%
CCCC(CCO)O	inf	26	100%	40	100%	15	87%	25	88%
CCCC=CCCCC	inf	47	100%	47	100%	26	96%	26	96%
CCCCCCCCCCCN	inf	7	100%	17	100%	4	50%	15	87%
CCCCCCCCOC(=O)C	inf	8	100%	29	100%	2	100%	11	100%
CCCCOP(O)O	inf	7	100%	8	100%	3	100%	4	100%
CCCCS	inf	10	100%	15	100%	2	50%	4	75%
CCCOC(c1ccccc1)	inf	36	100%	37	100%	14	100%	14	100%
CN=C	inf	18	100%	18	100%	4	75%	5	60%
CN1CCNCC1	inf	9	100%	14	100%	9	100%	11	100%
CNS(=O)(=O)c1ccc(cc1)	inf	5	100%	10	100%	2	100%	7	100%
COC(=O)c1ccc(cc1)C	inf	3	100%	6	100%	1	100%	1	100%
COP(=S)(O)	inf	13	100%	13	100%	4	100%	5	100%
CSc1ccc(cc1)	inf	8	100%	10	100%			1	100%
FCC(F)(F)	inf	10	100%	11	100%	2	100%	2	100%
n1sccc1	inf	12	100%	12	100%	7	100%	7	100%
NC(=O)Nc1ccccc1	inf	15	100%	15	100%	4	75%	6	83%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
Nc1ccccn1	inf	12	100%	14	100%	1	100%	1	100%
O=C(c1cccc2c1nccc2)O	inf	6	100%	7	100%	1	100%	1	100%
O=CCCC(=O)O	inf	13	100%	20	100%	7	86%	11	91%
OC(=O)CC(=O)	inf	4	100%	5	100%	2	100%	2	100%
Oc1ccc(cc1)C=C	inf	20	100%	21	100%	9	100%	10	100%
Oc1ccc2c(c1)cc(n2)	inf	3	100%	4	100%	2	100%	2	100%
OCC1OC(CC1O)n1ccc(nc1=O)	inf	8	100%	8	100%	4	100%	4	100%
OCCOC(=O)C=C	inf	8	100%	12	100%	2	100%	4	100%
S(=O)(=O)c1cccc2c1nccc2	inf	9	100%	9	100%	4	100%	4	100%
c1ccc(cc1Cl)Cl	10.18	31	97%	68	99%	17	71%	37	86%
CCSC	5.77	12	92%	39	97%	6	83%	17	88%
COC(C)(C)C	4.1	9	89%	28	96%	5	80%	9	78%
Oc1ccc(c(c1)Cl)	3.95	9	100%	27	96%			10	100%
OCCOCCOC	3.8	9	89%	26	96%	9	78%	18	83%
OCCN(CC)C	3.11	15	87%	43	95%	8	100%	20	95%
C(=O)OCCC	2.99	21	90%	124	95%	15	93%	56	96%
c1ccc(cc1C)Cl	2.89	10	90%	40	95%	6	83%	11	91%
Oc1ccc(cc1)C(=O)	2.35	18	94%	33	94%	5	80%	11	82%
Clc1ccc(cc1)	2.29	38	89%	177	94%	13	92%	71	89%
CCCCOc1ccc(cc1)	2.28	2	100%	16	94%			10	100%
OC(=O)CO	2.18	8	100%	46	93%	2	100%	16	100%
CC(CCC=C(C))C	2.05	9	78%	58	93%	7	100%	34	91%
COc1cccc(c1)C	1.63	19	95%	47	91%	11	82%	25	88%
NCCCN	1.59	10	90%	23	91%	5	100%	12	92%
COc1cccc(c1)C(=O)	1.52			11	91%			2	100%

		Training Set				Prediction Set			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
C(=O)c1cccc(c1[N])	1.44	7	100%	21	90%	2	100%	6	100%
Cc1ccc(cc1)Cl	1.4			41	90%			11	91%
C(=O)Nc1ccccc1	1.39	18	89%	71	90%	7	71%	35	77%
N#Cc1ccc(cc1)	1.22	7	100%	9	89%	1	100%	3	100%
OCc1ccc(cc1)Cl	1.22			18	89%			4	75%
CNCCCO	1.09	2	100%	49	88%	1	100%	23	91%
N#C	1.04	22	95%	55	87%	7	86%	15	73%

Table F. Total occurrences and percentage of correct assignments, for the FALSE NEGATIVE fragments extracted by SARpy from the TN-FN dataset. The results are shown for both training and prediction set, as evaluated using either the complete SARpy ruleset or the "Single Fragments" approach (each fragment considered separately).

		Training Set				**Prediction Set**			
		SARpy		Single Fragments		SARpy		Single Fragments	
Fragment	Training LR	tot	% correct	tot	% correct	tot	% correct	tot	% correct
c1nc(c([nH]1)c1ccccc1)c1ccccc1	inf	5	100%	5	100%	1	100%	1	100%
ONC(=O)c1ccccc1	inf	18	100%	18	100%	6	100%	6	100%
CC12OOC1c1c(O2)cccc1	52.67	9	89%	9	89%	2	50%	2	50%
COO	16.46	12	58%	21	71%	3	33%	5	40%
CC(=O)NO	9.88	10	60%	10	60%	4	50%	4	50%
[N](=O)	3.66	21	43%	28	36%	14	29%	22	27%
CC(=CC)c1ccc(cc1)	3.29	9	56%	15	33%	2	0%	4	0%
ClCCl	2.71	11	55%	24	29%	7	43%	9	33%
CCCl	2.44	10	50%	37	27%	2	0%	12	33%
COCCOC(=O)	1.98	5	60%	26	23%			6	0%
NCN	1.86	22	41%	50	22%	10	10%	27	19%
Cc1ccc(cc1)[N]	1.76	18	33%	38	21%	10	30%	19	26%
CC(CCC(O)(C)C)	1.54	3	67%	37	19%	4	0%	27	7%
CC1CCC(O1)	1.32	7	57%	30	17%	4	0%	16	13%
c1ccc(cc1)OC	1.23	30	23%	171	16%	13	15%	82	13%
c1ccc(cc1)[N]	1.21	35	34%	206	16%	25	20%	120	18%
C(Cc1ccccc1)C	1.21	14	21%	116	16%	11	0%	59	5%
CCOCC(CO)	1.1	3	100%	63	14%	1	0%	34	9%

Annex B.
Chemical classes

III. Chemical classes identified within the mutagenicity dataset

Here is reported the complete output of the istChemFeat software, relative to the 6065 molecules with experimental Ames test values used for the study. Using the SMILES representation of the molecules, the tool compared the dataset with a library of functional groups and atom-centered fragments. For each chemical classes, istChemFeat calculate the total number of matches within the dataset, and the distribution of the property. In this case, the number and percentage of mutagenic and non-mutagenic chemicals within each classes were calculated.

Table G. Chemical classes identified within the mutagenicity dataset. Within this study the chemical classes were defined by the presence of either a functional group (group no. N) or an atom-centered fragment (ACF). For each class are reported the total number of matches (molecules containing the group or ACF), the number and percentage of mutagens.

Group	Matches	Mutagens	Mutagens (%)
(group no. 1) terminal primary C(sp3)	3192	1569	49%
(group no. 2) total secondary C(sp3)	2435	1137	47%
(group no. 3) total tertiary C(sp3)	798	273	34%
(group no. 4) total quaternary C(sp3)	292	73	25%
(group no. 5) ring secondary C(sp3)	1282	680	53%
(group no. 6) ring tertiary C(sp3)	513	202	39%
(group no. 7) ring quaternary C(sp3)	199	49	25%
(group no. 8) aromatic C(sp2)	4350	2566	59%
(group no. 9) unsubstituted benzene C(sp2)	3973	2378	60%
(group no. 10) substituted benzene C(sp2)	4021	2387	59%
(group no. 11) non-aromatic conjugated C(sp2)	1789	965	54%
(group no. 12) terminal primary C(sp2)	280	109	39%
(group no. 13) aliphatic secondary C(sp2)	990	496	50%
(group no. 14) aliphatic tertiary C(sp2)	458	184	40%

Group	Matches	Mutagens	Mutagens (%)
(group no. 16) terminal C(sp)	13	5	38%
(group no. 17) non-terminal C(sp)	18	6	33%
(group no. 20) isocyanates (aliphatic)	3	0	0%
(group no. 21) isocyanates (aromatic)	5	3	60%
(group no. 22) thiocyanates (aliphatic)	2	0	0%
(group no. 23) thiocyanates (aromatic)	1	1	100%
(group no. 24) isothiocyanates (aliphatic)	2	1	50%
(group no. 25) isothiocyanates (aromatic)	3	2	67%
(group no. 26) carboxylic acids (aliphatic)	370	112	30%
(group no. 27) carboxylic acids (aromatic)	109	41	38%
(group no. 28) esters (aliphatic)	514	217	42%
(group no. 29) esters (aromatic)	141	36	26%
(group no. 30) primary amides (aliphatic)	30	14	47%
(group no. 31) primary amides (aromatic)	16	6	38%
(group no. 32) secondary amides (aliphatic)	321	153	48%
(group no. 33) secondary amides (aromatic)	82	55	67%
(group no. 34) tertiary amides (aliphatic)	81	20	25%
(group no. 35) tertiary amides (aromatic)	15	6	40%
(group no. 36) (thio-) carbamates (aliphatic)	57	40	70%
(group no. 37) (thio-) carbamates (aromatic)	23	10	43%
(group no. 38) acyl halogenides (aliphatic)	16	15	94%
(group no. 39) acyl halogenides (aromatic)	16	15	94%
(group no. 45) thioesters (aromatic)	1	0	0%
(group no. 48) aldehydes (aliphatic)	107	67	63%
(group no. 49) aldehydes (aromatic)	39	13	33%
(group no. 50) ketones (aliphatic)	248	98	40%
(group no. 51) ketones (aromatic)	421	260	62%
(group no. 52) urea (-thio) derivatives	153	71	46%
(group no. 53) carbonate (-thio) derivatives	8	0	0%
(group no. 54) amidine derivatives	36	15	42%
(group no. 55) guanidine derivatives	15	11	73%
(group no. 56) imines (aliphatic)	11	3	27%
(group no. 57) imines (aromatic)	12	3	25%
(group no. 58) oximes (aliphatic)	13	6	46%
(group no. 59) oximes (aromatic)	14	4	29%
(group no. 60) primary amines (aliphatic)	196	78	40%
(group no. 61) primary amines (aromatic)	604	442	73%
(group no. 62) secondary amines (aliphatic)	158	62	39%
(group no. 63) secondary amines (aromatic)	179	111	62%
(group no. 64) tertiary amines (aliphatic)	264	128	48%
(group no. 65) tertiary amines (aromatic)	197	114	58%
(group no. 66) N hydrazines	66	42	64%
(group no. 67) N azo-derivatives	123	83	67%
(group no. 68) nitriles (aliphatic)	76	27	36%
(group no. 69) nitriles (aromatic)	32	16	50%
(group no. 70) positively charged N	174	123	71%
(group no. 71) quaternary N	22	7	32%
(group no. 72) hydroxylamines (aliphatic)	65	53	82%
(group no. 73) hydroxylamines (aromatic)	94	70	74%
(group no. 74) N-nitroso groups (aliphatic)	190	171	90%
(group no. 75) N-nitroso groups (aromatic)	15	9	60%

Group	Matches	Mutagens	Mutagens (%)
(group no. 76) nitroso groups (aliphatic)	2	2	100%
(group no. 77) nitroso groups (aromatic)	46	41	89%
(group no. 78) nitro groups (aliphatic)	43	26	60%
(group no. 79) nitro groups (aromatic)	898	763	85%
(group no. 80) imides (-thio)	74	34	46%
(group no. 81) hydrazones	28	16	57%
(group no. 82) hydroxyl groups	1493	677	45%
(group no. 83) aromatic hydroxyls	596	287	48%
(group no. 84) primary alcohols	368	172	47%
(group no. 85) secondary alcohols	513	245	48%
(group no. 86) tertiary alcohols	115	53	46%
(group no. 87) ethers (aliphatic)	663	374	56%
(group no. 88) ethers (aromatic)	674	367	54%
(group no. 91) anhydrides (-thio)	12	1	8%
(group no. 93) thiols	20	7	35%
(group no. 95) sulfides	155	72	46%
(group no. 96) disulfides	21	8	38%
(group no. 97) sulfoxides	7	2	29%
(group no. 98) sulfones	71	47	66%
(group no. 100) sulfinic (thio-/dithio-) acids	2	0	0%
(group no. 101) sulfonic (thio-/dithio-) acids	53	17	32%
(group no. 102) sulfuric (thio-/dithio-) acids	13	6	46%
(group no. 103) sulfites (thio-/dithio-)	1	0	0%
(group no. 104) sulfonates (thio-/dithio-)	31	24	77%
(group no. 105) sulfates (thio-/dithio-)	4	4	100%
(group no. 106) sulfonamides (thio-/dithio-)	106	42	40%
(group no. 107) phosphites/thiophosphites	5	1	20%
(group no. 108) phosphates/thiophosphates	88	27	31%
(group no. 110) phosphonates (thio-)	12	5	42%
(group no. 112) CH2RX	307	223	73%
(group no. 113) CHR2X	60	41	68%
(group no. 114) CR3X	11	1	9%
(group no. 115) R=CHX	26	17	65%
(group no. 116) R=CRX	47	35	74%
(group no. 118) CHRX2	60	35	58%
(group no. 119) CR2X2	9	0	0%
(group no. 120) R=CX2	30	22	73%
(group no. 121) CRX3	84	29	35%
(group no. 122) X on aromatic ring	655	266	41%
(group no. 123) X on ring C(sp3)	38	14	37%
(group no. 124) X on ring C(sp2)	60	27	45%
(group no. 125) X on exo-conjugated C	104	60	58%
(group no. 126) Aziridines	50	48	96%
(group no. 127) Oxiranes	300	219	73%
(group no. 128) Thiranes	3	1	33%
(group no. 129) Azetidines	1	1	100%
(group no. 130) Oxetanes	6	4	67%
(group no. 132) Beta-Lactams	8	0	0%
(group no. 133) Pyrrolidines	42	11	26%
(group no. 134) Oxolanes	90	42	47%
(group no. 135) tetrahydro-thiophenes	2	0	0%

Group	Matches	Mutagens	Mutagens (%)
(group no. 136) Pyrroles	152	93	61%
(group no. 137) Pyrazoles	14	13	93%
(group no. 138) Imidazoles	205	157	77%
(group no. 139) Furanes	155	109	70%
(group no. 140) Thiophenes	54	43	80%
(group no. 141) Oxazoles	10	1	10%
(group no. 142) Isoxazoles	4	0	0%
(group no. 143) Thiazoles	81	58	72%
(group no. 144) Isothiazoles	84	52	62%
(group no. 145) Triazoles	18	4	22%
(group no. 146) Pyridines	508	308	61%
(group no. 147) Pyridazines	26	10	38%
(group no. 148) Pyrimidines	67	42	63%
(group no. 149) Pyrazines	84	61	73%
(group no. 150) 1-3-5-Triazines	12	3	25%
(group no. 152) donor atoms for H-bonds (N and O)	3244	1725	53%
(group no. 153) acceptor atoms for H-bonds (N,O,F)	5608	3073	55%
(C-001) CH3R / CH4	2602	1232	47%
(C-002) CH2R2	1700	694	41%
(C-003) CHR3	556	165	30%
(C-004) CR4	292	73	25%
(C-005) CH3X	1121	629	56%
(C-006) CH2RX	2175	1121	52%
(C-007) CH2X2	129	72	56%
(C-008) CHR2X	1300	674	52%
(C-009) CHRX2	280	155	55%
(C-010) CHX3	7	3	43%
(C-011) CR3X	378	155	41%
(C-012) CR2X2	89	36	40%
(C-013) CRX3	96	33	34%
(C-014) CX4	13	9	69%
(C-015) =CH2	227	72	32%
(C-016) =CHR	840	420	50%
(C-017) =CR2	459	184	40%
(C-018) =CHX	140	89	64%
(C-019) =CRX	220	108	49%
(C-020) =CX2	55	40	73%
(C-021) #CH	13	5	38%
(C-022) #CR / R=C=R	18	6	33%
(C-024) R--CH--R	4199	2497	59%
(C-025) R--CR--R	3442	2082	60%
(C-026) R--CX--R	2914	1712	59%
(C-027) R--CH--X	438	226	52%
(C-028) R--CR--X	755	503	67%
(C-029) R--CX--X	166	100	60%
(C-030) X--CH--X	33	26	79%
(C-031) X--CR--X	36	20	56%
(C-032) X--CX--X	30	9	30%
(C-033) R--CH..X	211	129	61%
(C-034) R--CR..X	507	341	67%
(C-035) R--CX..X	213	169	79%

Group	Matches	Mutagens	Mutagens (%)
(C-036) Al-CH=X	135	77	57%
(C-037) Ar-CH=X	81	43	53%
(C-038) Al-C(=X)-Al	273	111	41%
(C-039) Ar-C(=X)-R	230	111	48%
(C-040) R-C(=X)-X / R-C#X / X=C=X	1878	832	44%
(C-041) X-C(=X)-X	275	140	51%
(C-042) X--CH..X	89	56	63%
(C-043) X--CR..X	128	88	69%
(C-044) X--CX..X	140	107	76%
(H-046) H attached to C0(sp3) no X attached to next C	1984	904	46%
(H-047) H attached to C1(sp3)/C0(sp2)	5728	3199	56%
(H-048) H attached to C2(sp3)/C1(sp2)/C0(sp)	702	406	58%
(H-049) H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	770	422	55%
(H-050) H attached to heteroatom	3254	1727	53%
(H-051) H attached to alpha-C	1025	456	44%
(H-052) H attached to C0(sp3) with 1X attached to next C	1527	632	41%
(H-053) H attached to C0(sp3) with 2X attached to next C	298	168	56%
(H-054) H attached to C0(sp3) with 3X attached to next C	59	36	61%
(H-055) H attached to C0(sp3) with 4X attached to next C	1	1	100%
(O-056) alcohol	1069	530	50%
(O-057) phenol / enol / carboxyl OH	1042	432	41%
(O-058) #NOME?	2894	1426	49%
(O-059) Al-O-Al	870	492	57%
(O-060) Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X	1447	731	51%
(O-061) O--	958	799	83%
(O-062) O- (negatively charged)	54	27	50%
(O-063) R-O-O-R	41	31	76%
(Se-064) Any-Se-Any	3	3	100%
(N-066) Al-NH2	172	72	42%
(N-067) Al2-NH	150	56	37%
(N-068) Al3-N	260	128	49%
(N-069) Ar-NH2 / X-NH2	649	468	72%
(N-071) Ar-NAl2	183	110	60%
(N-072) RCO-N< / >N-X=X	1108	628	57%
(N-073) Ar2NH / Ar3N / Ar2N-Al / R..N..R	457	300	66%
(N-074) R#N / R=N-	235	110	47%
(N-075) R--N--R / R--N--X	988	615	62%
(N-076) Ar-NO2 / R--N(--R)--O / RO-NO	907	770	85%
(N-077) Al-NO2	75	58	77%
(N-078) Ar-N=X / X-N=X	397	326	82%
(N-079) N+ (positively charged)	146	109	75%
(F-081) F attached to C1(sp3)	7	2	29%
(F-082) F attached to C2(sp3)	13	0	0%
(F-083) F attached to C3(sp3)	63	12	19%
(F-084) F attached to C1(sp2)	101	65	64%
(F-085) F attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X	15	6	40%
(Cl-086) Cl attached to C1(sp3)	243	168	69%
(Cl-087) Cl attached to C2(sp3)	82	41	50%
(Cl-088) Cl attached to C3(sp3)	36	20	56%
(Cl-089) Cl attached to C1(sp2)	546	222	41%
(Cl-090) Cl attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X	126	80	63%

Group	Matches	Mutagens	Mutagens (%)
(Br-091) Br attached to C1(sp3)	128	88	69%
(Br-092) Br attached to C2(sp3)	19	14	74%
(Br-093) Br attached to C3(sp3)	5	3	60%
(Br-094) Br attached to C1(sp2)	96	39	41%
(Br-095) Br attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X	7	4	57%
(I-096) I attached to C1(sp3)	8	6	75%
(I-097) I attached to C2(sp3)	1	0	0%
(I-098) I attached to C3(sp3)	2	1	50%
(I-099) I attached to C1(sp2)	10	2	20%
(I-100) I attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X	1	1	100%
(S-106) R-SH	25	10	40%
(S-107) R2S / RS-SR	468	273	58%
(S-108) R=S	72	25	35%
(S-109) R-SO-R	7	2	29%
(S-110) R-SO2-R	196	75	38%
(Si-111) >Si<	14	4	29%
(B-112) >B- as in boranes	2	1	50%
(P-115) P ylids	1	0	0%
(P-117) X3-P=X (phosphate)	109	41	38%
(P-118) PX3 (phosphite)	5	1	20%
(P-119) PR3 (phosphine)	1	0	0%
(P-120) C-P(X)2=X (phosphonate)	14	4	29%

IV. Secondary chemical classes identified

The three tables below report the complete outputs of the istChemFeat software, relative to the three primary chemical classes used as case studies: nitro aromatics (898 molecules), aliphatic hydroxylamines (65 molecules) and aliphatic tertiary amides (81 molecules). Using the SMILES representation of the molecules, the tool compared the dataset with a library of functional groups and atom-centered fragments. For each chemical class, istChemFeat calculates the total number of matches within the dataset, and the distribution of the property. In this case, the number and percentage of mutagenic and non-mutagenic chemicals within each class were calculated.

Table H. Secondary chemical classes identified for nitro aromatic molecules.

Chemical Feature	Matches	Mutagens	Mutagens (%)
(group no. 1) terminal primary C(sp3)	287	222	77%
(group no. 2) total secondary C(sp3)	156	129	83%
(group no. 3) total tertiary C(sp3)	17	13	76%
(group no. 4) total quaternary C(sp3)	14	3	21%
(group no. 5) ring secondary C(sp3)	81	72	89%
(group no. 6) ring tertiary C(sp3)	5	3	60%
(group no. 7) ring quaternary C(sp3)	1	0	0%
(group no. 8) aromatic C(sp2)	898	763	85%
(group no. 9) unsubstituted benzene C(sp2)	799	669	84%
(group no. 10) substituted benzene C(sp2)	804	669	83%
(group no. 11) non-aromatic conjugated C(sp2)	261	238	91%
(group no. 12) terminal primary C(sp2)	6	6	100%
(group no. 13) aliphatic secondary C(sp2)	91	84	92%
(group no. 14) aliphatic tertiary C(sp2)	9	7	78%
(group no. 23) thiocyanates (aromatic)	1	1	100%
(group no. 25) isothiocyanates (aromatic)	2	1	50%
(group no. 26) carboxylic acids (aliphatic)	13	9	69%
(group no. 27) carboxylic acids (aromatic)	24	21	88%
(group no. 28) esters (aliphatic)	33	25	76%
(group no. 29) esters (aromatic)	18	16	89%
(group no. 30) primary amides (aliphatic)	2	2	100%
(group no. 31) primary amides (aromatic)	5	5	100%
(group no. 32) secondary amides (aliphatic)	39	34	87%
(group no. 33) secondary amides (aromatic)	27	27	100%
(group no. 34) tertiary amides (aliphatic)	3	3	100%
(group no. 35) tertiary amides (aromatic)	1	1	100%

Chemical Feature	Matches	Mutagens	Mutagens (%)
(group no. 36) (thio-) carbamates (aliphatic)	2	2	100%
(group no. 38) acyl halogenides (aliphatic)	4	4	100%
(group no. 39) acyl halogenides (aromatic)	4	4	100%
(group no. 48) aldehydes (aliphatic)	4	4	100%
(group no. 49) aldehydes (aromatic)	6	6	100%
(group no. 50) ketones (aliphatic)	5	1	20%
(group no. 51) ketones (aromatic)	38	32	84%
(group no. 52) urea (-thio) derivatives	12	11	92%
(group no. 54) amidine derivatives	1	1	100%
(group no. 56) imines (aliphatic)	2	0	0%
(group no. 57) imines (aromatic)	2	1	50%
(group no. 59) oximes (aromatic)	8	4	50%
(group no. 60) primary amines (aliphatic)	4	2	50%
(group no. 61) primary amines (aromatic)	98	87	89%
(group no. 62) secondary amines (aliphatic)	2	2	100%
(group no. 63) secondary amines (aromatic)	36	31	86%
(group no. 64) tertiary amines (aliphatic)	14	14	100%
(group no. 65) tertiary amines (aromatic)	29	29	100%
(group no. 66) N hydrazines	7	7	100%
(group no. 67) N azo-derivatives	21	18	86%
(group no. 68) nitriles (aliphatic)	9	9	100%
(group no. 69) nitriles (aromatic)	10	7	70%
(group no. 70) positively charged N	13	10	77%
(group no. 71) quaternary N	1	1	100%
(group no. 72) hydroxylamines (aliphatic)	4	4	100%
(group no. 73) hydroxylamines (aromatic)	7	5	71%
(group no. 74) N-nitroso groups (aliphatic)	4	4	100%
(group no. 75) N-nitroso groups (aromatic)	2	2	100%
(group no. 77) nitroso groups (aromatic)	6	5	83%
(group no. 78) nitro groups (aliphatic)	3	3	100%
(group no. 80) imides (-thio)	5	5	100%
(group no. 81) hydrazones	12	12	100%
(group no. 82) hydroxyl groups	114	82	72%
(group no. 83) aromatic hydroxyls	79	50	63%
(group no. 84) primary alcohols	31	25	81%
(group no. 85) secondary alcohols	33	23	70%
(group no. 87) ethers (aliphatic)	24	20	83%
(group no. 88) ethers (aromatic)	66	57	86%
(group no. 91) anhydrides (-thio)	1	1	100%
(group no. 95) sulfides	9	8	89%
(group no. 96) disulfides	5	5	100%
(group no. 98) sulfones	2	2	100%
(group no. 101) sulfonic (thio-/dithio-) acids	9	3	33%
(group no. 104) sulfonates (thio-/dithio-)	2	2	100%
(group no. 106) sulfonamides (thio-/dithio-)	23	20	87%
(group no. 108) phosphates/thiophosphates	5	4	80%
(group no. 110) phosphonates (thio-)	1	1	100%
(group no. 112) CH2RX	19	17	89%
(group no. 116) R=CRX	2	2	100%
(group no. 118) CHRX2	2	1	50%
(group no. 120) R=CX2	3	3	100%

Chemical Feature	Matches	Mutagens	Mutagens (%)
(group no. 121) CRX3	8	3	38%
(group no. 122) X on aromatic ring	96	74	77%
(group no. 125) X on exo-conjugated C	1	1	100%
(group no. 126) Aziridines	2	2	100%
(group no. 127) Oxiranes	10	9	90%
(group no. 134) Oxolanes	4	4	100%
(group no. 136) Pyrroles	39	36	92%
(group no. 137) Pyrazoles	5	5	100%
(group no. 138) Imidazoles	51	44	86%
(group no. 139) Furanes	84	84	100%
(group no. 140) Thiophenes	37	37	100%
(group no. 143) Thiazoles	32	31	97%
(group no. 144) Isothiazoles	19	19	100%
(group no. 146) Pyridines	36	34	94%
(group no. 148) Pyrimidines	13	13	100%
(group no. 149) Pyrazines	8	7	88%
(group no. 150) 1-3-5-Triazines	2	2	100%
(group no. 152) donor atoms for H-bonds (N and O)	395	326	83%
(group no. 153) acceptor atoms for H-bonds (N,O,F)	898	763	85%
(C-001) CH3R / CH4	231	179	77%
(C-002) CH2R2	112	93	83%
(C-003) CHR3	12	8	67%
(C-004) CR4	14	3	21%
(C-005) CH3X	105	91	87%
(C-006) CH2RX	168	140	83%
(C-007) CH2X2	7	5	71%
(C-008) CHR2X	58	43	74%
(C-009) CHRX2	14	12	86%
(C-011) CR3X	5	5	100%
(C-013) CRX3	8	3	38%
(C-015) =CH2	3	3	100%
(C-016) =CHR	87	81	93%
(C-017) =CR2	9	7	78%
(C-018) =CHX	2	2	100%
(C-019) =CRX	7	5	71%
(C-020) =CX2	6	4	67%
(C-024) R--CH--R	871	741	85%
(C-025) R--CR--R	629	532	85%
(C-026) R--CX--R	794	660	83%
(C-027) R--CH--X	44	38	86%
(C-028) R--CR--X	107	101	94%
(C-029) R--CX--X	24	23	96%
(C-030) X--CH--X	6	6	100%
(C-031) X--CR--X	11	11	100%
(C-032) X--CX--X	2	2	100%
(C-033) R--CH..X	66	66	100%
(C-034) R--CR..X	161	154	96%
(C-035) R--CX..X	146	142	97%
(C-036) Al-CH=X	5	5	100%
(C-037) Ar-CH=X	30	26	87%
(C-038) Al-C(=X)-Al	6	1	17%

Chemical Feature	Matches	Mutagens	Mutagens (%)
(C-039) Ar-C(=X)-R	24	19	79%
(C-040) R-C(=X)-X / R-C#X / X=C=X	213	187	88%
(C-041) X-C(=X)-X	14	13	93%
(C-042) X--CH..X	17	15	88%
(C-043) X--CR..X	31	26	84%
(C-044) X--CX..X	38	37	97%
(H-046) H attached to C0(sp3) no X attached to next C	193	148	77%
(H-047) H attached to C1(sp3)/C0(sp2)	892	761	85%
(H-048) H attached to C2(sp3)/C1(sp2)/C0(sp)	89	85	96%
(H-049) H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	101	89	88%
(H-050) H attached to heteroatom	396	327	83%
(H-051) H attached to alpha-C	78	60	77%
(H-052) H attached to C0(sp3) with 1X attached to next C	66	56	85%
(H-053) H attached to C0(sp3) with 2X attached to next C	18	18	100%
(H-054) H attached to C0(sp3) with 3X attached to next C	1	1	100%
(O-056) alcohol	75	52	69%
(O-057) phenol / enol / carboxyl OH	115	79	69%
(O-058) =O	288	248	86%
(O-059) Al-O-Al	45	35	78%
(O-060) Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X	183	164	90%
(O-061) O--	898	763	85%
(O-062) O- (negatively charged)	10	7	70%
(N-066) Al-NH2	4	2	50%
(N-067) Al2-NH	2	2	100%
(N-068) Al3-N	14	14	100%
(N-069) Ar-NH2 / X-NH2	102	91	89%
(N-071) Ar-NAl2	29	29	100%
(N-072) RCO-N< / >N-X=X	126	119	94%
(N-073) Ar2NH / Ar3N / Ar2N-Al / R..N..R	105	92	88%
(N-074) R#N / R=N-	51	44	86%
(N-075) R--N--R / R--N--X	150	139	93%
(N-076) Ar-NO2 / R--N(--R)--O / RO-NO	898	763	85%
(N-077) Al-NO2	7	7	100%
(N-078) Ar-N=X / X-N=X	38	31	82%
(N-079) N+ (positively charged)	10	7	70%
(F-083) F attached to C3(sp3)	8	3	38%
(F-084) F attached to C1(sp2)	11	10	91%
(F-085) F attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X	1	1	100%
(Cl-086) Cl attached to C1(sp3)	13	13	100%
(Cl-087) Cl attached to C2(sp3)	3	2	67%
(Cl-089) Cl attached to C1(sp2)	68	50	74%
(Cl-090) Cl attached to C2(sp2)-C4(sp2)/C1(sp)/C4(sp3)/X	13	13	100%
(Br-091) Br attached to C1(sp3)	6	4	67%
(Br-094) Br attached to C1(sp2)	11	11	100%
(I-099) I attached to C1(sp2)	4	1	25%
(S-106) R-SH	1	1	100%
(S-107) R2S / RS-SR	121	119	98%
(S-108) R=S	5	3	60%
(S-110) R-SO2-R	17	8	47%
(P-117) X3-P=X (phosphate)	6	4	67%
(P-120) C-P(X)2=X (phosphonate)	1	1	100%

Table I. Secondary chemical classes identified for aliphatic hydroxylamines.

Chemical Feature	Matches	Mutagens	Mutagens (%)
(group no. 1) terminal primary C(sp3)	38	33	87%
(group no. 2) total secondary C(sp3)	24	19	79%
(group no. 3) total tertiary C(sp3)	7	2	29%
(group no. 4) total quaternary C(sp3)	3	3	100%
(group no. 5) ring secondary C(sp3)	9	4	44%
(group no. 6) ring tertiary C(sp3)	7	2	29%
(group no. 8) aromatic C(sp2)	44	42	95%
(group no. 9) unsubstituted benzene C(sp2)	43	41	95%
(group no. 10) substituted benzene C(sp2)	43	41	95%
(group no. 11) non-aromatic conjugated C(sp2)	41	40	98%
(group no. 13) aliphatic secondary C(sp2)	4	3	75%
(group no. 14) aliphatic tertiary C(sp2)	1	0	0%
(group no. 26) carboxylic acids (aliphatic)	3	0	0%
(group no. 30) primary amides (aliphatic)	1	0	0%
(group no. 32) secondary amides (aliphatic)	1	0	0%
(group no. 33) secondary amides (aromatic)	1	1	100%
(group no. 36) (thio-) carbamates (aliphatic)	5	4	80%
(group no. 49) aldehydes (aromatic)	1	1	100%
(group no. 52) urea (-thio) derivatives	11	7	64%
(group no. 54) amidine derivatives	1	1	100%
(group no. 60) primary amines (aliphatic)	2	2	100%
(group no. 62) secondary amines (aliphatic)	2	2	100%
(group no. 64) tertiary amines (aliphatic)	1	1	100%
(group no. 65) tertiary amines (aromatic)	2	2	100%
(group no. 69) nitriles (aromatic)	1	1	100%
(group no. 77) nitroso groups (aromatic)	1	1	100%
(group no. 79) nitro groups (aromatic)	4	4	100%
(group no. 82) hydroxyl groups	25	16	64%
(group no. 83) aromatic hydroxyls	2	2	100%
(group no. 85) secondary alcohols	1	1	100%
(group no. 88) ethers (aromatic)	6	6	100%
(group no. 106) sulfonamides (thio-/dithio-)	1	0	0%
(group no. 122) X on aromatic ring	9	8	89%
(group no. 133) Pyrrolidines	3	0	0%
(group no. 138) Imidazoles	1	1	100%
(group no. 146) Pyridines	2	1	50%
(group no. 152) donor atoms for H-bonds (N and O)	33	21	64%
(group no. 153) acceptor atoms for H-bonds (N,O,F)	65	53	82%
(C-001) CH3R / CH4	37	32	86%
(C-002) CH2R2	23	18	78%
(C-003) CHR3	3	0	0%
(C-004) CR4	3	3	100%
(C-005) CH3X	16	11	69%
(C-006) CH2RX	36	35	97%
(C-008) CHR2X	5	5	100%
(C-011) CR3X	7	2	29%
(C-016) =CHR	3	2	67%

Chemical Feature	Matches	Mutagens	Mutagens (%)
(C-017) =CR2	1	0	0%
(C-019) =CRX	1	1	100%
(C-020) =CX2	3	3	100%
(C-024) R--CH--R	43	41	95%
(C-025) R--CR--R	40	39	98%
(C-026) R--CX--R	22	20	91%
(C-027) R--CH--X	1	0	0%
(C-028) R--CR--X	3	2	67%
(C-033) R--CH..X	1	1	100%
(C-037) Ar-CH=X	1	1	100%
(C-040) R-C(=X)-X / R-C#X / X=C=X	51	46	90%
(C-041) X-C(=X)-X	13	9	69%
(C-042) X--CH..X	1	1	100%
(H-046) H attached to C0(sp3) no X attached to next C	18	16	89%
(H-047) H attached to C1(sp3)/C0(sp2)	57	49	86%
(H-048) H attached to C2(sp3)/C1(sp2)/C0(sp)	1	1	100%
(H-049) H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	4	3	75%
(H-050) H attached to heteroatom	33	21	64%
(H-051) H attached to alpha-C	33	29	88%
(H-052) H attached to C0(sp3) with 1X attached to next C	24	19	79%
(H-053) H attached to C0(sp3) with 2X attached to next C	3	3	100%
(O-056) alcohol	25	16	64%
(O-057) phenol / enol / carboxyl OH	5	2	40%
(O-058) =O	60	50	83%
(O-059) Al-O-Al	35	33	94%
(O-060) Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X	38	37	97%
(O-061) O--	4	4	100%
(N-066) Al-NH2	2	2	100%
(N-067) Al2-NH	1	1	100%
(N-068) Al3-N	1	1	100%
(N-069) Ar-NH2 / X-NH2	2	1	50%
(N-071) Ar-NAl2	2	2	100%
(N-072) RCO-N< / >N-X=X	26	19	73%
(N-073) Ar2NH / Ar3N / Ar2N-Al / R..N..R	1	1	100%
(N-074) R#N / R=N-	2	2	100%
(N-075) R--N--R / R--N--X	3	2	67%
(N-076) Ar-NO2 / R--N(--R)--O / RO-NO	4	4	100%
(N-077) Al-NO2	32	32	100%
(N-078) Ar-N=X / X-N=X	2	2	100%
(F-084) F attached to C1(sp2)	2	2	100%
(Cl-089) Cl attached to C1(sp2)	4	4	100%
(Br-094) Br attached to C1(sp2)	3	2	67%
(S-110) R-SO2-R	1	0	0%

Table J. Secondary chemical classes identified for aliphatic tertiary amides.

Chemical Feature	Matches	Mutagens	Mutagens (%)
(group no. 1) terminal primary C(sp3)	54	12	22%
(group no. 2) total secondary C(sp3)	65	14	22%
(group no. 3) total tertiary C(sp3)	22	3	14%

Chemical Feature	Matches	Mutagens	Mutagens (%)
(group no. 5) ring secondary C(sp3)	47	9	19%
(group no. 6) ring tertiary C(sp3)	15	3	20%
(group no. 8) aromatic C(sp2)	65	15	23%
(group no. 9) unsubstituted benzene C(sp2)	55	13	24%
(group no. 10) substituted benzene C(sp2)	57	13	23%
(group no. 11) non-aromatic conjugated C(sp2)	24	8	33%
(group no. 13) aliphatic secondary C(sp2)	18	5	28%
(group no. 26) carboxylic acids (aliphatic)	16	1	6%
(group no. 32) secondary amides (aliphatic)	17	1	6%
(group no. 64) tertiary amines (aliphatic)	8	0	0%
(group no. 82) hydroxyl groups	31	2	6%
(group no. 85) secondary alcohols	8	0	0%
(group no. 87) ethers (aliphatic)	10	2	20%
(group no. 88) ethers (aromatic)	9	3	33%
(group no. 95) sulfides	14	1	7%
(group no. 112) CH2RX	11	5	45%
(group no. 122) X on aromatic ring	9	1	11%
(group no. 132) Beta-Lactams	8	0	0%
(group no. 144) Isothiazoles	8	3	38%
(group no. 152) donor atoms for H-bonds (N and O)	41	5	12%
(group no. 153) acceptor atoms for H-bonds (N,O,F)	81	20	25%
(C-001) CH3R / CH4	46	10	22%
(C-002) CH2R2	54	11	20%
(C-003) CHR3	14	3	21%
(C-005) CH3X	43	9	21%
(C-006) CH2RX	63	15	24%
(C-007) CH2X2	8	2	25%
(C-008) CHR2X	42	6	14%
(C-009) CHRX2	14	1	7%
(C-011) CR3X	10	1	10%
(C-016) =CHR	12	2	17%
(C-019) =CRX	8	3	38%
(C-024) R--CH--R	59	14	24%
(C-025) R--CR--R	54	11	20%
(C-026) R--CX--R	32	9	28%
(C-027) R--CH--X	8	1	13%
(C-028) R--CR--X	9	3	33%
(C-035) R--CX..X	10	5	50%
(C-040) R-C(=X)-X / R-C#X / X=C=X	79	19	24%
(H-046) H attached to C0(sp3) no X attached to next C	24	4	17%
(H-047) H attached to C1(sp3)/C0(sp2)	81	20	25%
(H-048) H attached to C2(sp3)/C1(sp2)/C0(sp)	26	4	15%
(H-049) H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	13	3	23%
(H-050) H attached to heteroatom	41	5	12%
(H-051) H attached to alpha-C	47	12	26%
(H-052) H attached to C0(sp3) with 1X attached to next C	47	10	21%
(H-053) H attached to C0(sp3) with 2X attached to next C	9	1	11%
(O-056) alcohol	14	0	0%
(O-057) phenol / enol / carboxyl OH	18	3	17%
(O-058) =O	81	20	25%
(O-059) Al-O-Al	13	4	31%

Chemical Feature	Matches	Mutagens	Mutagens (%)
(O-060) Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X	16	6	38%
(N-068) Al3-N	8	0	0%
(N-072) RCO-N< / >N-X=X	81	20	25%
(N-075) R--N--R / R--N--X	18	4	22%
(Cl-086) Cl attached to C1(sp3)	9	4	44%
(Cl-089) Cl attached to C1(sp2)	8	2	25%
(S-107) R2S / RS-SR	27	5	19%